JISC Final Report

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	Dryad-UK		
Project Hashtag	#dryaduk		
Start Date	October 1st 2010	**End Date**	October 4th 2011
Lead Institution	Oxford University		
Project Director	David Shotton		
Project Manager	Brian Hole		
Contact email	david.shotton@zoo.ox.ac.uk, brian.hole@ubiquitypress.com		
Partner Institutions	The British Library, DCC, Charles Beagrie Ltd.		
Project Web URL	http://datadryad.org/		
Programme Name	Managing Research Data (JISCMRD)		
Programme Manager	Simon Hodson		

Document Information			
Author(s)	Brian Hole and David Shotton		
Project Role(s)	Project Manager and Project Director		
Date	19 Jan 2012	**Filename**	finalreportDryadUK_1-1.doc
URL	http://wiki.datadryad.org/DryadUK		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
0.1	06.09.11	Initial creation of document (BH)
0.2	04.10.11	Final draft of document with BL content (BH)
0.3	12.10.11	Oxford University content added (DS)
0.4	17.10.11	DCC content added (AW)
0.5	18-10-2011	Revised (BH and DS)
0.6	20.10.11	Revised and added appendices (AW)
1.0	21.12.11	Added recommendations, document URLs, workshop appendices and minor revisions (BH)
1.1	19.01.12	Final revisions incorporating edits from DS and TV (BH)

Table of Contents

1 Acknowledgements

Dryad-UK was funded under the JISC Managing Research Data programme. We would like to acknowledge the support of the DCC (Angus Whyte and Kevin Ashley) and Charles Beagrie Ltd (Neal Beagrie) as project partners, as well as the members of the advisory board: Ed Pentz, Ian Handel, Michael Jubb, Theo Bloom and Peter Murray-Rust. We are also grateful to Todd Vision and his colleagues, particularly Ryan Scherle and Peggy Schaeffer, for assisting integration with the US Dryad activities. We also thank the many publishers who have contributed to the Dryad-UK Project and have 'come on board' Dryad during the past year.

2 Project Summary

The data behind research publications in the biosciences are not commonly made openly available to readers of those publications. The Dryad Data Repository, which only hosts data relating to peer-reviewed journal articles, solves this problem by integrating the open archiving of research datasets by authors with the journal article submission workflows of publishers. The Dryad-UK project has helped to extend Dryad in five main areas.

- By gathering feedback from publishers, funders and researchers in the UK, it has informed financial modelling and the sustainability planning for Dryad as an international organisation.
- It has increased the value of the repository for the UK research community by extending it to include 6 UK-based publishers and over 20 new journals, including those in the high impact areas of biomedicine and infectious diseases.
- In addition to this, by piloting a mirror installation of the repository at the British Library, the project has gathered data on a technical model for establishing a stable and scalable international infrastructure for Dryad in future.
- Since one concern regarding Dryad holdings is the paucity of their metadata, and since one focus for new Dryad-associated journals has been those dealing with infectious disease, the Dryad-UK project has been active in developing MIIDI, a Minimal Information standard for reporting an Infectious Disease Investigation, and tools to create MIIDI-compliant metadata.
- It has provided a framework for assessment of Dryad, by identifying important aspects to stakeholders and recommending indicators that, with further development, could be used for ongoing evaluation. The project also contributes further evidence on a potential benefit from depositing data; impact on citation rates for associated research articles

3 Main Body of Report

3.1 Project Outputs and Outcomes

Output / Outcome Type	Brief Description and URLs (where applicable)
Sustainability Planning Report	A report based on the draft report from Charles Beagrie in 2010, and expanded to include the new pricing plans, proposed governance structure, and other cost-projection information from the Dryad-UK project. This report will be completed in Q1 2012, and will then be available on the Dryad wiki at http://wiki.datadryad.org/DryadUK.
British Library workshop and report	The first Dryad-UK Workshop was held at the British Library on April 1st 2011. The aim of the workshop was to gather stakeholder feedback on appropriate funding models for Dryad as an international organisation. This involved examining mixed funding models involving subscriptions, submission fees, hosting of services, and grants, and participation from journals, publishers, authors, funding bodies and research institutions. A report summarising the outcomes of the workshop is at http://wiki.datadryad.org/DryadUK.

Oxford University workshop and presentations	The second Dryad-UK workshop, with the title **Research Data Sharing and the Dryad Repository** was held at Oxford University on September 12th 2011, attended by researchers, data managers and publishers. The presentations given at the workshop are available at: http://wiki.datadryad.org/DryadUK.
General presentations	Aspects of the Dryad-UK project were presented at a range of conferences during the year. These presentations are listed at: http://wiki.datadryad.org/DryadUK.
Publications	Two papers are in draft form and will be submitted for publication in late 2011. The first [1] is on Dryad's potential impact on citation rates to associated articles, from Heather Piwowar, and the second [2] is based on the community discussions on repository pricing models from the Dryad-UK BL workshop and Dryad Consortium Board meeting, from Brian Hole. These will be available at the following URL: http://wiki.datadryad.org/DryadUK.
Cooperation with new publishers	We approached the major UK-based academic publishers and formed relationships, whereby they were invited to contribute journals to Dryad, and gave feedback on the payment and governance plans of Dryad. This was particularly valuable, and has resulted in those plans being reformulated to reflect the needs of these UK stakeholders. These publishers either have now joined, or are in the process of joining Dryad and integrating applicable journals from their catalogues. The publishers involved are: • BiomedCentral • The BMJ • The British Ecological Society • Elsevier • Oxford University Press • PLoS • Wiley Blackwell The integration status of journals from these publishers can be found in the Dryad wiki at: http://wiki.datadryad.org/Journal_Integration_Status.
Integration of new journals in the biosciences	• BMC Ecology • BMC Evolutionary Biology • Animal Behaviour (Elsevier) • Behavioural Processes (Elsevier) • Ecological Modelling (Elsevier) • Biological Conservation (Elsevier) • Biomaterials (Elsevier) • Gene (Elsevier) • General and Comparative Endocrinology (Elsevier) • Genomics (Elsevier) • Comparative Biochemistry & Physiology Part D – Genomics & Proteomics (Elsevier) • Gene Expression Patterns (Elsevier) • PLoS Biology • PLoS Computational Biology • PLoS Genetics • Systematic Biology (OUP) • (Wiley and other OUP titles not yet confirmed)
Integration of new journals in biomedicine and	• BMJ Open • Infection, Genetics & Evolution (Elsevier) • Toxicon (Elsevier)

infectious disease	• PLoS Medicine • PLoS Neglected Tropical Diseases • PLoS One • PLoS Pathogens • (Wiley and OUP titles not yet confirmed)
The Dryad mirror at the British Library	The mirror of the Dryad Data Repository installed at the British Library has provided the following data: • Proof of concept data that file and database replication mechanisms are feasible. • Proof of concept of redirection of European visitors to the UK mirror. • Proof of concept of the failover mechanism ensuring that Dryad is fully available should one of the mirror nodes fail.
Mapping the DataCite Metadata Kernel to RDF	Since Dryad uses DataCite DOIs, we have mapped the DataCite Metadata Kernel version 2.0 to RDF, to facilitate the publication of metadata accompanying Dryad datasets to RDF. Mapping document at http://bit.ly/nGrhln.
Exemplar mappings of Dryad metadata to RDF	The DataCite2RDF mapping was then used to create exemplar mappings of the metadata accompanying a Dryad Data Package and the Dryad Data Files it contained to RDF. The RDF output is at http://bit.ly/qSkUyc.
The MIIDI Metadata Standard, editor and input form	MIIDI, a Minimal Information standard for reporting an Infectious Disease Investigation, was proposed in 2009 as a method of defining appropriate metadata to accompany a journal article or a data set relating to an infectious disease investigation. During the Dryad-UK Project, this has now been developed into a formal XML schema. The MIIDI Web site, giving information about MIBBI and the development activities surrounding it, has been established at http://www.miidi.org. A software system, the MIIDI Metadata Editor, has been created that uses the MIIDI schema to create a MIIDI Metadata Input Form for recording MIIDI-compliant metadata, validates the entered metadata as a MIIDI Report, and permits its export in a number of other formats including XHTML/RDFa, JSON and various serializations of RDF. The MIIDI Metadata Input Form is a Web form created from the MIIDI schema by the MIIDI Metadata Editor, that permits convenient entry of MIIDI-compliant metadata. It can be accessed at http://www.miidi.org:8080/input-form/. Blog posts on the MIIDI work are available on the Open Citations and Semantic Publishing Blog at http://opencitations.wordpress.com. In addition to permitting rich metadata to be created concerning an infectious disease dataset, the MIIDI system also permits metadata to be recorded about the related journal article. Such metadata can be thought of as a structured digital summary of the paper, that can be separately published as an *Open Research Report in* (*name of disease)*. The vision of Open Research Reports was presented at the conference Science Online London 2011, and is available at http://imageweb.zoo.ox.ac.uk/pub/2011/presentations/Shotton-ScienceOnlineLondon2011-OpenResearchReports.pdf.
A Framework for Assessing the Dryad	Provides assessment criteria and indicators that maybe used to assess Dryad. The report identifies stakeholders and how they were consulted on the

Data Repository (report)	criteria. It also summarises how DryadUK contributed on those criteria that relate to its aims. http://wiki.datadryad.org/DryadUK.
Dryad data deposit report	Provides an overview of Dryad data deposit rates from journals, considering their integration status. The report will be available in Q1 2012 at: http://wiki.datadryad.org/DryadUK.

3.2 How the outputs / outcomes were achieved

Dryad journal membership, funding model and sustainability

One of the major aims of Dryad-UK has been to inform sustainability plans for Dryad as an international organisation. To this end, we organised a Dryad-UK workshop at the British Library on April 1st 2011 and invited a range of stakeholders from the publisher, journal, researcher and funding communities, to debate the appropriateness of the proposed funding and governance models (see Workshop Report: http://wiki.datadryad.org/DryadUK). This workshop provided detailed and constructive feedback on how the existing models needed to be changed if they were to satisfy the requirements of the UK community, especially with regard to the larger publishers. This feedback was then used to reformulate the plans at the Dryad Executive Committee meeting in Vancouver in July 2011. On the following days, the Dryad Consortium Board, including several representatives from UK-based publishers (The BMJ, PLoS and Wiley Blackwell), discussed the plans and reworked them further until a consensus was reached that was workable for the majority of parties involved, including those from the UK. These plans were then further reviewed by members of the Dryad-UK project who had experience in forming international non-profit organisations (Adam Farquhar, Ed Pentz and Neil Beagrie), and the results fed back to Dryad in the Sustainability Planning Report (to be completed in Q1 2012, and will then be available at: http://wiki.datadryad.org/DryadUK).

Dryad-UK also sought to extend the involvement of Dryad with UK-based publishers and journals. Initial conversations held by the British Library with the larger publishers did not immediately result in positive feedback. We therefore continued to engage with the publishers through the first Workshop held at the British Library in April, and through a series of further meetings during the year. By actively engaging with the publishers and incorporating their feedback into the reworking of the payment and governance plans, we were able to secure the involvement of six of the major publishers, along with the initial integration of over 20 relevant journal titles.

Dryad server infrastucture

During the Dryad-UK Project, and partly as a result of it, Dryad has expanded over the past year, both in becoming an internationally governed organisation and in terms of the steadily increasing volume of its content from a wider range of international sources (see Dryad data deposition report at, to be completed in Q2 2011 and then available at: http://wiki.datadryad.org/DryadUK). Because of this, its infrastructure needs to be able to scale correspondingly. To assist in this, we set up a mirror server at the British Library in January 2011, and engaged in a pilot study to test the technology required to improve the scalability, stability, security and performance of the Dryad infrastructure through a distributed architecture. The principle focus of the pilot was that of validating the use of the simplest and most cost-effective solutions to replication. This was achieved through the automated copying of files, and of the entire Dryad database, on a scheduled basis. Redirection of users located in Europe to the UK mirror was also successfully tested, but planned failover testing could not be completed in the project timeframe.

Piloting approaches to metadata enhancement: the MIIDI Metadata Standard, Editor and Input Form

The MIIDI Metadata Standard, which prior to the Dryad-UK project was only outlined as a set of textual terms, has now been propertly defined as an XML data model. In addition, many (although not yet all) its terms have been mapped to RDF using appropriate ontologies, and these metadata terms have been included in the data model as RDFa attributes. This XML data model now defines the developing MIIDI Metadata Standard, available from the MIIDI web site at http://www.miidi.org.

An application, the MIIDI Metadata Editor, has been developed using a combination of Open Source software (XForms and Orbeon Forms) and W3C standard languages. Its purpose is to transform an annotated XML Schema file into an input form. In this way, the MIIDI XML data model can be used to create a MIIDI Metadata Input Form that is displayed to the user in a Web browser (see http://www.miidi.org:8080/input-form/).

The MIIDI Input Form allows the creation of a MIIDI Report that is validated for content and structure against the original XML Schema file.

The purpose of the MIIDI Metadata Editor is to facilitate the creation of rich human- and machine-readable metadata describing an infectious disease investigation, and the research outputs from this investigation in terms of datasets and journal articles, software, mathematical models and experimental workflows.

To make the task of metadata entry easier, the Editor software includes the following functionalities:

- The ability to save a partially completed MIIDI Report as a customised template.
- The ability to upload an existing partially completed MIIDI Report into the MIIDI Form to permit editing and further metadata entry.
- The ability to validate input values against the underlying XML schema, flagging datatype inconsistencies such as incorrect date formats.
- An auto-completion service, which will return full bibliographic metadata from PubMed upon submission of a valid PubMed ID or DOI.
- A look-up service for ontology terms from defined biomedical ontologies, using the BioPortal API.
- The ability to capture numerical geo-coordinates from a selected location using Google Maps.
- The ability to save the completed MIIDI Report as a validated XML file.
- The ability to view and export the MIIDI Report in HTML, in JSON format, and in a number of RDF serializations.

While being developed to work with MIIDI for the annotation of infectious disease datasets for submission to Dryad, we have been aware that the MIIDI Metadata Editor application has the potential to be re-purposed for other metadata-gathering purposes, for example for assisting users to record metadata complying with other MIBBI standards, simply by substituting the relevant domain-specific components within the generic data model. We are thus working to create such a generic editor.

Assessment and Evaluation

A further aim in Dryad-UK was to better understand how to evaluate the usefulness of Dryad data publication to the scientific community, and particularly UK journals and publishers. The evaluation aimed to deliver an “assessment framework for future independent evaluations”. This draws on three usually distinct types of evaluation [3]. Formative assessment has a project’s process as its focus, aiming to improve a project’s ability to achieve its outputs or outcomes. Later ‘summative’ assessment is normally of those outputs or outcomes, and ‘impact analysis’ investigates their longer-term consequences. Combining these strands involved consultation on appropriate criteria to express

Dryad's value to stakeholders, and consideration of the indicators or metrics that would be needed to track whether Dryad provides the value sought from it and achieves longer-term impacts.

Criteria were drafted by the DCC following a review of Dryad documentation, which articulates the value of the data repository from the project's perspective, and from repository evaluation practice [4]. The two stakeholder workshops described earlier were the main focus for consultation. The first of these included breakout sessions on the 'value of Dryad', notes from which were used to refine the list of criteria. Questionnaires were also used to obtain ratings on the importance of the criteria. This was repeated with a second questionnaire and shorter list at the second workshop, which included a small number of researchers. An online survey of Dryad depositors and reusers was also carried out but with very low response, probably due to its timing in the vacation period.

The assessment report proposes metrics that reflect established measures for assessing repositories (e.g. on trust) and other online resources (e.g. on usability or access). They include indicators of user satisfaction that would be gathered from questionnaires, plus measures of Dryad's activity and policies. These may be assessed on publicly available information that should also be expected to be available from alternative repositories.

The evaluation strand of the project will submit for publication a study [1] reviewing and contributing to the evidence of a 'citation benefit' from data deposit, from higher citation rates to articles that openly accessible datasets relate to. In a previous study Heather Piwowar and colleagues found a large increase in the citation rate of publications where data had been made publicly available [5]. However the study was relatively small, and unable to control for potential confounding factors. This larger analysis, of publications in the same domain (microarray analysis of gene expression in human clinical studies), gives a more robust and complete picture of the relationship between data sharing and citation rates.

As part of this work package, we also provide an end-of-project summation of the UK project's contribution.

Lastly the 'Planning Report' (to be completed in Q1 2012 and then available at: http://wiki.datadryad.org/DryadUK) describes the rate of new submissions to Dryad. NESCent monitored the rate of data submissions from journals, relative to the number of articles published. Relating this data to the timing of journal policy changes allowed an assessment of the impact of these on deposit rates.

3.3 Lessons learned

The project resulted in key learnings in five areas: sustainability and governance, expansion, technology, metadata, and evaluation.

Sustainability and governance

The project provided valuable feedback on the sustainability and governance of Dryad. A workshop was held at the British Library on April 1st, 2011, attended by representatives of publishers, journals, societies, researchers and funding bodies. The aim of the workshop was to gather stakeholder feedback on appropriate funding models for Dryad as an international organisation, and examined proposed mixed funding models involving subscriptions, submission fees, hosting of services, and grants. Participants were also asked about the value of Dryad to the community, and the role that partner institutions such as the British Library should play in it. Overall, the value of Dryad to the community was seen as very high, and the role of supporting institutions as valuable. However, larger UK-based publishers expressed concerns about the preliminary pricing models, which they did not feel were scalable. A full summary of the workshop is available online (http://wiki.datadryad.org/DryadUK).

The feedback from the meeting was used by the Dryad Executive over the following months to formulate new pricing plans that would satisfy the requirements of the larger publishers. These plans were then presented to Dryad members and stakeholders at the Dryad Consortium Board meeting in

Vancouver in July, including representatives of larger UK publishers from the Dryad-UK project (PLoS, Wiley and the BMJ). Once again there was a large amount of debate, which led to the proposals being further reworked until they were flexible enough to satisfy the majority of stakeholders. Further information on these outcomes is available in the Dryad-UK Sustainability Planning Report (to be completed in Q1 2012, and will be available at: http://wiki.datadryad.org/DryadUK).

Expansion

The project was successful in expanding the range of UK-based publishers and journals integrated with Dryad to a large degree, because of the above-mentioned ongoing contact and engagement. Because of this, larger publishers such as Elsevier, OUP and Wiley were given confidence that Dryad was responsive to their requirements, and would be a system that would not only meet their journals' and authors' requirements but would also be affordable to operate.

In order to enable integration the larger publishers such as BiomedCentral, BMJ and PLoS, it was necessary to adapt to their specific requirements, resulting in several new features on the Dryad platform:

- Integration with new editorial management systems: experience was gained with the Editorial Manager system at PLoS, the ScholarOne system at the BMJ, and with BMC's proprietary system, all of which will make it easier to integrate with other publishers in future.
- Enabling peer review of data: It is now possible for journals to request that data be deposited with Dryad at the time of article submission, rather than at the time of article publication, as hitherto. This enables reviewers to look at the submitted data while peer-reviewing the paper, resulting in higher quality of peer review. This is achieved by keeping the relevant datasets private in Dryad, while providing a password to the journal so that editors and peer reviewers can access the data if required. This was an important feature for the UK publishers, with the BMJ Open for example promoting it publicly.
- Early provision of DOIs: Provisional DOIs for datasets can now also be created at the time of submission, so that the journals have sufficient time to include correct data references within the related articles.
- Delaying publication of data: While metadata regarding new records in Dryad had formerly been made public on the date of article acceptance, journals now have the option of embargoing release of metadata until the date the article itself is published. This is an important feature for a few publishers who protect details of forthcoming research from advance publicity.

Technology

A mirror of the Dryad server was set up at the British Library and piloted, with the aim of understanding the technology required and cost involved. This was successful in demonstrating that a simple process of copying files and data on a scheduled or event-driven basis, as well as location-based user redirection worked well. The cost of hosting and maintaining the server was also tracked (see Dryad-UK Sustainability Planning Report, to be completed in Q1 2012 and will be available at: http://wiki.datadryad.org/DryadUK). Not all of the project goals were met here due to technical resource constraints at the North Carolina State University Digital Library, the host of Dryad in the USA. While relatively straightforward to implement, failover testing was not completed, and there was also not time to compare response times for UK-based users with and without the mirror and redirection. However, it is hoped that these tests will be carried out in Q1 of 2012.

Staffing

The lessons that we have learned with regard to staffing are that short JISC projects run by small research groups lacking permanent staff are at the mercy of the job market. Skilled personnel for this sort of work are in short supply, and it is difficult to recruit such people to short-term contracts on University salaries – they either want the job security of a longer period of employment, or an 'industrial' salary to compensate for the insecurity of a short contract.

In the Dryad-UK project, the University of Oxford advertised the post immediately after receiving the award letter, and succeeded in appointing a well-qualified person. For reasons due to commitments to his previous employment, the start date was delayed at his request until 11th November. Shortly before taking up the post, he informed us he had decided to stay with his previous employer, because he had been offered a new job there at ~£20K per year more that we could pay. Our post was then re-advertised, and after interviews was offered to a second applicant, to start in early January 2011, following a visit home to see his family in India over Christmas. From India, he contacted us to say that he would be unable to take up his post, because he now had to care for his sick father. A third advertisement enabled us to secure the services of Tanya Gray, who started work on 30 January 2011, four months after the project start date.

Fortunately, this saga has had a happy ending, since Tanya has brought to the project exceptional skills in XML software design and metadata management that surpass those of the previous appointees. However, these delays have meant that the Oxford work relating to the use of MIIDI tools to provide rich metadata to accompany the submission of datasets relating to infectious disease investigations to Dryad is as yet incomplete. We have submitted a separate proposal to the JISC Programme Manager for continuation of this work over the next few months, as outlined under Future Impact below.

Assessment and Evaluation

The Framework for Assessment lists sixteen criteria in three groups; 'quality of interaction', 'take-up and impact', and 'policy and process'. In drafting the criteria, the main sources were literature on Dryad, including its public wiki and the case presented for funding. These articulate the project team's view of Dryad's value. Stakeholders views on the draft criteria were also obtained through workshop discussion and questionnaires.

Participants in the first workshop rated an initial draft list of twelve criteria. Criteria were added and removed based on these responses and the discussion notes, then a second iteration was rated at the final workshop, and the few online survey responses received were taken into account. Criteria were included if they were rated 'very important' by most of the respondents to at least one of the questionnaires.

The Assessment Framework report describes how each of the criteria shown in table 1 below relates to DryadUK project outputs and Dryad's capabilities more generally. The report describes what was learned from engaging with DryadUK stakeholders on criteria relevant to the project, as summarised in the Appendices, and outlines Dryad's current position on other criteria. Proposed indicators are summarised in the later section titled 'Future Impact'.

Assessment criteria	*Criterion is of relevance to the work of Dryad-UK*
1. Interaction Quality	
1.1. Deposit a wide range of data types	✓
1.2. Deposit process usable	✓
1.3. Data subject to peer review	✓
1.4. Discoverable	✓
1.5. Machine readable	✓
2. Take-up and Impact	
2.1. Access stats available	✓
2.2. Citable and attributable	✓
2.3. Data/ article citation impacts traceable	✓
2.4. Visible through repository interoperability	✓
2.5. Evidence of community take-up available	✓
2.6. Seen as best practice exemplar	

3. Policy and Process	
3.1. Open access licence terms apply	
3.2. Embargo period options given	
3.3. 'Trusted Digital Repository' status	
3.4. Clear curation service levels	
3.5. Representative governance	✓

Table 1 Assessment criteria for DryadUK

3.4 Immediate Impact

The project has had an immediate impact in several areas: with UK-based publishers and journals, with UK and international researchers, with the Dryad Community, at the British Library, and at the University of Oxford. We have also worked with the wider academic and publishing community by creating best practice guidelines for citing data, and by providing ontologies and mappings to assist DataCite and Dryad metadata to be published as open linked data.

UK-based publishers and journals
The project has helped UK publishers and journals both to find a solution to the problem of supplementary data, and also to raise their profiles in the scholarly community. With UK-specific modifications to Dryad, the journals are now also available to make available data to article reviewers for peer review. The value of this to publishers was highlighted by the BMJ in their summer newsletter:

> *BMJ Open* is the first medical journal to partner with Dryad-UK, an online repository, run by staff based at the British Library and University of Oxford, which provides a permanent, citable, and accessible home for datasets related to peer reviewed published articles in biosciences.
>
> Data sharing aims to help scientists and doctors validate and scrutinise researchers' findings in a bid to prevent fraud and eradicate the kind of selective reporting that has enabled some treatments to acquire regulatory approval, based on incomplete and biased data.
>
> In some cases this lack of transparency has prompted the subsequent restriction or withdrawal of certain treatments because of patient safety or effectiveness concerns, which were already evident in the unpublished data.
>
> The next step for the Group may be to mandate data deposition as a pre-requisite for peer review.

By being fully involved in the process of determining Dryad's pricing and governance plans, the UK publishers have also been able to immediately affect the way the Dryad system works, ensuring that it best meets their needs. This is evidenced by the new Sustainability Plan, which reflects a good deal of UK feedback.

Additionally, the Dryad-UK workshops have helped to stoke debate in the UK about open data and repository funding.

UK and international researchers
The Dryad-UK project has also initiated a process whereby a significant amount of data associated with articles in UK publications will be made openly available. This should lead to increased citation and collaboration for UK researchers, and greater demonstrable impact for the 2014 REF.

The wider Dryad community
Due to the active input of the UK publishers in determining the new pricing and governance plans through Dryad-UK, the wider Dryad community has also benefited from a more robust and scalable plan for Dryad, which is likely to provide a stable and sustainable future for the organisation and its members. Dryad colleagues in the USA have also benefitted from their interactions with those on the Dryad-UK Project in a variety of ways, not least in helping to strengthen Dryad's international presence, reputation and contacts.

The British Library
Through the Dryad-UK project, the British Library has come to better understand publisher and researcher requirements with regard to data, and to achieved a better understanding of what the wider community believes the Library's role should be. This has led to ongoing internal discussions about the possible future role of the library in holding data and partnering with other organisations such as Dryad. The project has also been an excellent demonstrator project for DataCite at the BL, helping them to demonstrate the value of data DOIs to publishers. It has also helped the BL Datasets Programme to increase its interaction with a broader range of communities.

University of Oxford
The University of Oxford's role in the Dryad-UK Project has been three-fold: First, it catalysed UK interest in the Dryad Data Repository and saw the potential for its internationalization. Second, it acted as the lead institution for the funding application and the funded project, although the bulk of the hard work was in fact undertaken by the other partners, as originally planned. Third, it has used the project funding to expand the vision of Dryad in terms of rich machine-readable metadata to accompany datasets. Despite delays due to staffing problems discussed elsewhere in this report, the work accomplished has been very significant, moving a vision towards a reality. All the tools have now been developed to permit rich MIIDI-compliant metadata to be created for infectious disease datasets, and, by potential extension, for infectious disease journal articles and for datasets and articles in other research areas. The project has also permitted significant work to be undertaken on the following three items, the wider implications of which will be increasingly seen in the data publishing community.

Best practice for data citation
The project has helped to promote the correct citation of datasets within journal articles. Following the publication of a discussion paper entitled Data Citation Best Practice Discussion Document, Dr Shotton was invited to contribute substantially to the Data Publishing Policies and Guidelines document providing advice to authors of Pensoft Journals, an on-line Open Access publisher of biological journals that has recently become a Dryad partner.

Mapping the DataCite Metadata Kernel to RDF
One aim of the Dryad-UK Project has been to facilitate the publication of metadata accompanying Dryad datasets to RDF, enabling these metadata to become part of the web of open linked data, so they can be understood programmatically and integrated automatically with similar data from elsewhere.

Since Dryad uses DataCite DOIs, and DataCite has made a recommendation concerning metadata to accompany a dataset, we have mapped the DataCite Metadata Kernel version 2.0 to RDF. The DataCite Metadata Kernel specifies the minimal metadata, and optional metadata, that should accompany a DataCite DOI for the identification of a published data entity. Within the Metadata Kernel document there is an XML mapping of these metadata terms, using DCMI Metadata Terms, and an example encoded in XML.

With Silvio Peroni, David Shotton has thus published a mapping of the DataCite metadata elements to RDF using ontology terms from commonly used vocabularies, supplemented by terms from the SPAR (Semantic Publishing and Referencing) ontologies CiTO, CiTO4Data and FaBiO, and from a new DataCite Ontology that we created to provide four object properties lacking in other ontologies.

Creating exemplar mappings of Dryad metadata to RDF
Using the DataCite2RDF mapping, we then published as Google docs both an RDF mapping of the DataCite XML example, and an RDF mapping of the metadata for a Dryad repository holding, showing how DataCite2RDF can be used for real data.

Contributing to the evidence-base on data sharing impacts
There is substantial community interest in links between data deposit and subsequent rates of citations to related articles, given that citation counts may affect researchers' funding and career advancement. The results from our extended analysis of the citation rate of articles with open data are likely to be of considerable interest to researchers, publishers, funders, and institutions, and contribute to the case for archiving 'article-related' datasets.

Future Impact

The Dryad Data Repository
Having undertaken the sustainability study, catalyzed a revision of the financial charging model for Dryad services, enabled increased data submissions from the 20 new journals recently brought on-board, and assisted in the 'internationalization' of Dryad, the Dryad-UK Project will have an ongoing beneficial effect on the Dryad Data Repository *per se*, which is now the best recognised general biomedical data repository in the world.

The British Library and Dryad
The British Library involvement in the Dryad-UK project was an excellent opportunity for the Library to gain an understanding of the mechanisms, skills and costs involved in data archiving, and to explore its role both in supporting the publishing and scholarly communities and in potentially integrating data into its own collections in future.

At the end of the 12 months allocated to Dryad-UK, the Library is in principle interested in continuing to maintain the Dryad mirror server and in exploring ways to continue supporting Dryad and UK publishers, although no definite decisions have yet been taken. There are several areas in which the library could contribute valuable expertise to Dryad, for example in the areas of curation and long-term preservation, and it is hoped that further project funding may be found in the near future to enable such collaboration to take place. One area in which the library can definitely continue to support Dryad is as the UK operating agent for DataCite, by advising on the provision of DOIs to datasets according to DataCite regulations and best practices.

UK publishers and academia
The Dryad-UK project can be expected to have a lasting future impact for UK publishers and researchers, since the project has already led to Dryad becoming an established data repository for journal articles from six of the major UK publishers (BiomedCentral, BMJ, Elsevier, OUP, PLoS and Wiley). It is therefore likely that Dryad will have continuing involvement with these publishers, and that this influence will spread to other major UK-based academic publishers. Dryad-UK has also served as a useful model for other UK projects interfacing with publisher workflows.

Given that publishing data leads to increased citation of the related journal articles [5], and that UK academics are more likely to publish in UK-based journals than journals based elsewhere, this association between Dryad and the major UK academic publishers will have ongoing benefits for UK-based researchers, by

- promoting international dissemination of UK research, potentially leading to new international collaborations;
- increased the reuse of UK research data;
- improving the citation of data and articles published by UK authors; and
- increasing the standing of their departments and institutions in the forthcoming REF.

With over 20 new journals integrating with Dryad, the project has succeeded in laying the groundwork for the open archiving of a significant fraction of the datasets published annually by UK life science researchers.

Facilitating correct citation of archived research datasets, and enabling their metadata to participate in the web of open linked data
All the tools are now in place, and demonstrated by published exemplars, to permit proper citation of archived datasets in the reference lists of journal articles, and to enable the metadata describing these archived research datasets to participate in the web of open linked data. At present, data citation is messy and incomplete, and available metadata about archived datasets is virtually non-existent. Within five years we expect the situation to have improved dramatically, not least as a result of the efforts undertaken during the Dryad-UK Project.

Facilitating independent assessment of Dryad
The assessment framework proposes indicators for tracking the impacts of Dryad of importance to each stakeholder group, based on three main sources of information described below. The framework is a working draft, and intended to evolve.There is also a need to ensure that the indicators represent information that can realistically be gathered, both for Dryad and for repositories comparable with Dryad. These indicators and sources are summarised in the Appendices and described more fully in the assessment framework report.

Further work to enable rich metadata to be associated with research project outputs, including datasets and journal articles
Permission from the JISC has been granted to use unspent salary budget for the continued employment of Tanya Gray on Work Package 5 (Metadata standards) until 30 April 2012. This no-cost extension of the Dryad-UK project is enabling additional development work on the MIIDI metadata creation tools:

- Development of the **MIIDI Metadata Standard** as an XML schema.
- Development of a web form, the **MIIDI Metadata Input Form**, that permits convenient entry of MIIDI-compliant metadata.
- Development of a software system, the **MIIDI Metadata Editor,** that uses the MIIDI schema to create the MIIDI Metadata Input Form used to record metadata, validates the entered metadata as a **MIIDI Report**, and permits its export in a number of other formats including XHTML/RDFa, JSON and various serializations of RDF.

As of January 2012, the major items of progress on this work can be summarized as follows:

Enhancements to the MIIDI Metadata Editor and MIIDI Metadata Input Form

Major revision undertaken to the Web-based Metadata Input Form, to the structure and appearance of the various output formats of the completed MIIDI Reports, and to the functionality of the underlying MIIDI Editor. The XML data model of the MIIDI Standard is now held in versioned form in a GitHub repository at https://github.com/miidi/miidi/tree/master/xsd. The revised Web-based MIIDI Editor is available for use at http://www.miidi.org:8080/input-form/. An alternative Java-based metadata input system is also being tested: http://www.miidi.org/wiki/index.php/Java_MIIDI_metadata_creation_tool.

Unit tests still have to be developed for the software.

Documentation

Documentation has now been written to cover:

- architecture/design documentation
- support documentation (user manual)
- software installation instructions

Still to do: Documentation about computer-based training, performance assessment, release notes and unit tests. In addition, the MIIDI Web site/wiki at http://www.miidi.org has been thoroughly revised and updated.

User testing

Hitherto, David Shotton acted as surrogate test user for the project. Outside test users have now been invited to test and comment upon the system, for final revisions prior to registering MIIDI as an official MIBBI Standard.

Using the existing MIIDI Metadata Editor software to develop generic metadata services

We have removed the infection-specific metadata elements from the MIIDI model to create a draft generic research investigation metadata model – Generic Minimal Information for a Research Investigation (GeMIRI), to which new specific metadata elements meeting the requirements of other domains can be added. The MIIDI Editor system was successfully repurposed for cancer research during the Semantic Web Applications and Tools for Life Sciences Hackathon (http://www.ukoln.ac.uk/events/devcsi/life-sciences-hackdays/programme/index.html).

Open Research Reports

David Shotton and colleagues have recently proposed the vision of Open Research Reports (http://imageweb.zoo.ox.ac.uk/pub/2011/presentations/Shotton-ScienceOnlineLondon2011-OpenResearchReports.pdf), an initiative that aims to publish MIIDI-compliant structured digital abstracts, authored by domain experts, summarizing the key factual and rhetorical information contained within disease-related publications. An exemplar Open Research Report has been created and is available at http://imageweb.zoo.ox.ac.uk/pub/2012/OpenResearchReports/, and various actions are being undertaken to move ORR from vision to reality.

4 Conclusions

4.1 General conclusions

1. Dryad is seen as a respectable repository for academic research data.
2. Publishers recognise Dryad as involving a community effort, and because of this are willing to partner with it.
3. Dryad needs more international partners like the British Library.
4. We have succeeded in expanding the coverage of Dryad from a narrow area of biology (ecology and evolution) to a broader biological and biomedical scope.
5. Both subscription access and open access academic publishers will collaborate with data repositories to facilitate archiving and publication of related datasets if financial arrangements are appropriate.
6. Engaging new publishers to work with Dryad requires investment of time and effort.

4.2 Conclusions relevant to the wider community

7. If researchers would like to ensure the preservation of their research datasets, they should publish in journals linked to Dryad, or lobby their journal editors to partner with Dryad.
8. DataCite DOIs are available for the unique identification of research datasets, and provide a mechanism for their citation, gaining authors academic credit.
9. Researchers and editors should take on board the best practice recommendations, e.g. from Pensoft Journals, concerning how data citations should properly be constructed and deployed, in a manner that resembles as closely as possible the citation of journal articles.
10. Easy-to-use tools such as the MIIDI Metadata Editor are needed for creation of rich domain-specific metadata to accompany either datasets or papers.
11. Appropriate ontologies and mappings exist for representing metadata about data repository holdings in RDF, so that they may become part of the web of linked data, facilitating automated semantically-aware resource discovery.

12. The community currently lacks agreed assessment frameworks for data repositories. The project has proposed one for Dryad that is intended to help align repository development with stakeholder values.

4.3 Conclusions relevant to JISC

13. It is very difficult to complete an ambitious body of work within a one-year project.
14. It is very difficult to appoint well-qualified developers and research assistants in the data management field for one-year projects at academic (rather than commercial) salaries.
15. Consequently, funding of longer projects would be beneficial.

    These last three points were presented in expanded form in the Final Report of the JISC ADMIRAL Project (q.v.)

5 Recommendations

5.1 General recommendations

- Dryad needs to provide a range of pricing options that fit the needs of journals, publishers and authors. Deposition needs to be affordable both for an author publishing in a non-member journal, and for a large publisher with hundreds of journals.
- Dryad should continue to consult openly with its stakeholder community, and to ensure that its pricing is transparent in order to retain trust.
- Dryad should continue to explore the provision of enriched metadata. Support is needed for further development of domain-independent technological standards for capturing domain-specific content standards, as demonstrated in this project for MIIDI. Partnership with innovative publishers will be important to the practical implementation of this data.
- Dryad and DataCite should use the mapping of the metadata kernel to RDF to expose that metadata as open linked data.
- The assessment framework created as part of Dryad-UK should be used to guide Dryad's near-term reporting strategy. For the future, the assessment framework would benefit from more in-depth analysis of depositor needs, search interface usability and retrieval performance, citation impacts, guidance on studies of data citation, and acceptable cost-benefit trade-offs (see section 6.3).
- As Dryad grows, it needs to develop greater efficiency and technical ability in the areas of curation and long-term digital preservation. As such it should explore running projects in these areas with the British Library or a similar institution that is able to contribute expertise.

5.2 Recommendations for the wider community

- Stakeholder feedback during the Dryad-UK project clearly demonstrates that there is a definite need for repositories such as Dryad in all scholarly disciplines, and that this is supported by important trends in publishing (e.g. publishers less inclined to hold supplementary material, growing use of data peer review). The Dryad-UK project has also demonstrated that sustainable models for such repositories can be found, and that publishers are ready and willing to support them. Communities outside of the bio-medical area should consider whether they want to also be involved in Dryad, or to follow the model by establishing similar repositories where they are needed.
- Good practice in data citation is essential to the success of repositories such as Dryad, and is also in the best interests of authors and publishers. The recommendations in the Data Citation Best Practice Discussion Document should be considered, and the community should agree on, and facilitate, best practice in the near future.

- The British Library and other large public institutions should support Dryad and other similar endeavours where possible, as this is a role that the scholarly and publishing communities believe is firmly in line with their missions.
- Publishers and journals in the UK (and elsewhere) should take an active role in contributing experience, guidance and leadership to Dryad. Dryad is adopting a governance model controlled by its stakeholders, and active engagement in this mechanism is the best way to ensure that the model works to the benefit of each of these.

5.3 Recommendations for JISC

- It is difficult to source appropriately skilled technical staff for shorter JISC projects, and to retain them for the full duration of the project. If JISC were able to help with sourcing staff, either by maintaining a pool of suitable developers etc., or a database of suitably skilled resources, this could help. JISC support of DevCSI is already a very worthwhile step in this direction.
- JISC should look to fund UK projects that further build upon the lessons and best practices established by the Dryad-UK project. This could include establishment of similar repositories or integrating them with publisher workflows, improving data citation, improving the richness of metadata (perhaps adapting MIIDI for other purposes), or exposing metadata as linked data.

6 Implications for the future

6.1 Dryad

It is an incontrovertible fact that the Dryad Digital Repository has been strengthened as a result of the Dryad-UK Project, and that valuable international links have been made between the Dryad staff in North Carolina and those engaged in working for Dryad in the UK. We are thus grateful to the JISC for the generous funding that made this possible.

Of course, we are concerned for the future, since there is no additional funding to support the continuation of Dryad-UK activities, the US arm of Dryad is still financially dependent upon NSF grant funding, and sustainability from deposit fees is still untested. Nevertheless, the activities of the Dryad-UK project have greatly assisted in developing a sustainability plan that is viable in principle, and the whole Dryad endeavour sits well within the more general movement for the publication of research data that is central in the minds of governments, national academies such as the Royal Society, funding agencies, publishers and increasingly academics – see, for example [6]-[8].

6.2 MIIDI and MIIDI tools - ongoing work

Despite the excellent progress made on the MIIDI Metadata Editor, this software suite is still under development, and it has yet to be used ‘in anger’ to create rich metadata to accompany datasets submitted to Dryad – its original objective.

To further enhance the quality of the application and make it suitable for third-party use, it is necessary to complete a number of tasks that have not yet been possible due to the relatively short duration of Tanya Gray’s involvement in the Dryad-UK project. These are detailed in a separate report previously sent to the JISC Programme Manager, and are summarized here.

Improvements to MIIDI software

The MIIDI metadata input form
The functionality of the MIIDI Metadata Input Form (http://www.miidi.org:8080/input-form/) has been evolving, and is ‘almost there’. However, some final revisions remain to be undertaken and tested to achieve the required initial functionality. As part of this, we intend to undertake testing of the software by developing a series of automated unit tests.

Software review
At present, the MIIDI Metadata Editor uses an open source software package called Orbeon Forms (http://www.orbeon.com). Because of the frequent use of AJAX calls between client and server, we fear that the system's performance may not scale well when we have multiple simultaneous users. This needs to be tested. Because of this, we also wish to evaluate a new software package called BetterForm (http://www.betterform.de) that peforms a similar function to Orbeon Forms, but that for technical reasons might offer an alternative implementation for the MIIDI Input Form that would provide better performance under load.

Documentation
We intend to further develop the documentation for the MIIDI Metadata Editor software, to assist third-party users.

User testing
We wish to employ external test users, themselves domain experts in infectious diseases, to test the functionality and usability of the form, to develop metadata both for journal articles and datasets. This work needs to be undertaken with some urgency. Their feedback will be used by the developer to further improve the MIIDI system.

Registering MIIDI as a MIBBI standard
To date, because MIIDI has been under active development, we have held off from registering MIIDI as a metadata standard as part of MIBBI (Minimal Information for Biological and Biomedical Investigations; http://www.mibbi.org), an umbrella organization for the registration of minimal information metadata standards in the biomedical domain. As soon as our final specifications have been implemented, we will do this.

Using MIIDI as a basis for a more generic metadata specification
The MIIDI Metadata Standard includes both generic components concerning provenance information, the research investigation and project outputs (papers, datasets, etc), and domain-specific information concerning infectious diseases. We now wish to see whether, by enriching these generic elements with other elements common to other metadata standards we could devise a 'core' metadata model that could then be customized for new domains.

Using the MIIDI Metadata Editor to develop metadata services for other purposes
The MIIDI Metadata Editor software generates a metadata input form and associated functions directly from a single XML Schema file that defines a set of metadata requirements. No bespoke HTML encoding is required. There is potential for this application to be re-purposed, such that a new metadata entry form could be created simply by substituting a new metadata definition. However, this would require some further effort to produce and document a fully generic instance of the MIIDI Metadata Editor software for third-party use.

Open Research Reports
David Shotton and colleagues have recently proposed the vision of Open Research Reports, an initiative that aims to publish MIIDI-compliant structured summaries of the key factual and rhetorical information contained within infectious disease journal articles, and to publish these in both human-readable 'instant journal' format and in machine-readable format for use by those in the developing world unable to afford subscription access to the relevant journals. Such an Open Research Report can be created by a domain expert using the MIIDI Metadata Editor, enabling that tool to be used for a different but related purpose - to annotate journal articles rather than research datasets.

By summarizing key facts about papers in a standard manner, it is hoped that such Open Research Reports also serve a separate purpose, greatly simplifying the task of the authors of systematic reviews in correctly identifying the articles reporting clinical trials that fit the acceptance criteria for the review.

The full implementation of the Open Research Reports vision is clearly much greater than can be accomplished without specific funding, but we hope in the near future to create proof-of-concept exemplar reports to use when applying for such funding.

6.3 Further development of the assessment framework

A framework for assessment has been produced comprising identified stakeholder groups, criteria important to them, and indicators that may be used to compare Dryad against alternatives or track performance over time. Since this was a relatively small part of a short project, wider engagement is still needed to build a more comprehensive picture of the factors that matter most to the communities that data repositories need to engage with.

Specific suggestions for further work in this area are:

Wider study of what is important to data depositors. The project engaged with a small number of authors as depositors or reusers. A larger and more representative study of what they see as the important attributes of a data repository would benefit the publishing and data repository communities. Such a study would adapt the questionnaire trialled in DryadUK, supplemented with interviews. The support of publishing industry bodies such as STM (International Association of Scientific, Technical & Medical Publishers) or the ALPSP (Association of Learned and Professional Society Publishers) would be valuable in organising this.

Search interface usability and retrieval performance: The framework has emphasized usability in the deposit process as this is a priority for Dryad, and the DryadUK project did not impact on search usability directly. This is an obvious gap that needs to be filled if the framework is to be used for comparative evaluations. Relevance metrics to test retrieval performance are a further gap, and it should be possible to extend the framework to monitor the impact on search precision and recall of enriched metadata (e.g. from MIIDI).

Citation impacts: It should be feasible with in the next few years to carry out an analysis of the citation rates of articles in Dryad partner journals that have data deposited in Dryad, compared to those that are from the same journals but do not. In the longer term, it should be feasible to compare the impact of depositing in Dryad relative to other options (journal supplementary materials, researcher-managed websites, institutional repositories, or specialized repositories).

Guidance on studies of data citation: the project has not addressed the needs of intermediaries. For example subject librarians who provide support to faculty in tracking the citation impact of published articles might benefit from guidance on tracking data citations and their needs in this areas should be investigated.

Acceptable cost-benefit tradeoffs: the assessment framework identifies benefits but gives no guidance on the trade-offs between benefit and cost that are acceptable to users and other stakeholders. Methods to explore willingness to pay for varying levels of service have been tried and tested in the digital library community. These could usefully be applied to guide repositories on the level of investment they should make in particular goals, such as achieving certification as a trusted digital repository.

7 References

[1] Piwowar, H. ‘Carrots for public data archiving: further evidence of association between availability and citation rate’ (to be submitted)

[2] Hole, B. ‘Finding a sustainable model for data archiving: feedback from the DryadUK project’ (to be submitted)

[3] JISC (2007) Six Steps to Effective Evaluation: A handbook for programme and project managers. Available at: http://www.jisc.ac.uk/media/documents/programmes/digitisation/SixStepsHandbook.pdf

[4] Shearer K. Institutional repositories: towards the identification of critical success factors. Canadian journal of information and library science. 2003;27(3):89–108 Available at: http://hdl.handle.net/1880/381

[5] Piwowar HA, Day RS, Fridsma DB (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE* **2**(3): e308. doi:10.1371/journal.pone.0000308.

[6] Science as a Public Enterprise. The Royal Society's policy study and call for evidence. Particularly concerned with data publication as a means of sharing scientific information with the public.

[7] Geoffrey Boulton, Michael Rawlins, Patrick Vallance, Mark Walport (2011). Science as a public enterprise: the case for open data. *The Lancet* **377** (Issue 9778): 1633 - 1635, 14 May 2011. doi:10.1016/S0140-6736(11)60647-8.

[8] Making Open Data Real: A Public Consultation. A policy document and call for contributions from the Cabinet Office of the UK Government.

The following references are cited in the Appendices:

[9] Dasgupta A, Ghosh A, Kumar R, Olston C, Pandey S, Tomkins A. The discoverability of the web [Internet]. In: Proceedings WWW2007. ACM Press; 2007 [cited 2011 Sep 22]. p. 421.Available from: http://www.www2007.org/papers/paper592.pdf

[10] Dasdan A, Tsioutsiouliklis K, Velipasaoglu E. Web search engine metrics [Internet]. In: Proceedings WWW2010. ACM Press; 2010 [cited 2011 Sep 9]. p. 1343.Available from: http://dasdan.net/ali/publications.php.

[11] Rumsey S. A case analysis of registering research activity for institutional benefit. International Journal of Information Management. 2010;30(2):174–9. doi: 10.1016/j.ijinfomgt.2009.12.004. Available at: http://ora.ox.ac.uk/objects/uuid%3Ad71f378e-9a58-44fe-98c7-9d9eda9b0174

[12] RADAR: Researching a Data Asset Registry: Available at: http://radar.blogs.edina.ac.uk/.

[13] Shotton D. DataCite2RDF – Mapping DataCite Metadata Scheme Terms to ontologies | JISC Open Citations. Available from: http://opencitations.wordpress.com/2011/06/30/datacite2rdf-mapping-datacite-metadata-scheme-terms-to-ontologies-2/.

[11] DataONE 'Data Packaging'. Available at: http://mule1.dataone.org/ArchitectureDocs-current/design/DataPackage.html

[10] Shotton D. How to cite data Available from: http://opencitations.wordpress.com/2011/06/30/how-to-cite-data/

[11] Dryad wiki 'Harvesting Technology'. Available at: https://www.nescent.org/wg_dryad/Harvesting_Technology.

All Dryad-UK reports and outputs either are or will be available on the Dryad wiki: http://wiki.datadryad.org/DryadUK.

8 Appendices

8.1 Appendix A: Agenda for first workshop

DryadUK Sustainability Workshop

April 1st 2011, 1pm-4:30pm, at the British Library, 96 Euston Road, London

Facilitator: Kevin Ashley, Director of the Digital Curation Centre

Schedule

12:30pm Registration and coffee
1:00pm Workshop begins
2:30pm 20 min break
4:30pm Drinks and networking

Workshop aim

The aim of the workshop is to gather stakeholder feedback on appropriate funding models for Dryad as an international organisation. This will involve examining mixed funding models involving subscriptions, submission fees, hosting of services, and grants. We are therefore inviting representatives of publishers, journals, researchers, research institutions and funding bodies to attend.

Workshop themes

What is the value of Dryad for your community?
1. As researchers?
2. As publishers?
3. As journals?
4. As funding bodies?

What should the funding model be?
1. Per paper charges dominant
2. Annual subscriptions dominant
3. Per paper charges and subscriptions balanced

What role should funding bodies play?
1. Funding the repository
2. Funding the authors to deposit
3. Funding the publishers to participate

Discussion format and rules

The workshop will begin with two short background presentations, outlining the history of repository funding to date, and the development of Dryad to the present. We will then shift into breakout groups to discuss the above themes. Following the break, the points that emerge for this will be elaborated in a facilitated open discussion. Because we want everyone to have a say and for the conversation to be open and frank, “Chatham House Rules” will apply, whereby: “… participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s)… may be revealed.” We will thus ensure that all quotes etc. used following the workshop are anonymised.

8.2 Appendix B: Agenda for second workshop

DryadUK workshop: **Data Sharing and the Dryad Data Repository**

Wolfson College, Oxford, September 12th 2011

08:45	**Coffee and welcome**
09:00-9:30	**Introduction** David Shotton (Oxford University) Todd Vision (Dryad)
9:30-11:00	**1. Journals and data publishing** Tim Stevenson (BiomedCentral) Theo Bloom (PLoS) David Tempest (Elsevier) Richard Sands (BMJ Open)
11:00-11:30	**Coffee break**
11:30-13:00	**2. Meta data, data citation and credit** David Shotton (Oxford University) Max Wilkinson (DataCite) Diane Cabell (OeRC/Creative Commons)
13:00-14:00	**Lunch**
14:00-15:30 `	**3. Researcher needs and the repository landscape** Ross Mounce (University of Bath) Sebastian Shimeld (Oxford University) Mark Thorley (NERC) Lyubo Penev (Pensoft)
15:30-16:00	**Coffee break**
16:00-17:30	**4. Sharing infectious disease data** Catherine Moyes (Oxford University) Ian Handel (Edinburgh University) David Shotton (Oxford University) Jenny Molloy (Oxford University)
17:30	**Drinks and networking**
18:00-19:00	**DryadUK Advisory Board Meeting**
19:00	**Preprandial drinks for those attending dinner**
19:30	**Dinner**

8.3 Appendix C: Developing the assessment framework

[Note the following is an extract from the Assessment Framework Report]

Two stakeholder workshops were held during the project and notes from the discussions have informed this report. The first of these was held at the British Library in April 2011. This was aimed primarily at groups of funders, publishers, and journal editors (Society-led and others) and included breakout sessions on the 'value of Dryad'. The second workshop, at Oxford University in September 2011, also included a small number of researchers. The workshops involved a people who could be considered 'early adopters' of Dryad and these were consulted using questionnaires on the relative importance of the draft criteria. These were made available in both face-to-face workshops and also online. Notes from the workshop discussions highlighting salient points on Dryad's value to the participants were also used as a 'reality check' on the draft criteria.

Most of the questionnaire responses came from the workshops: 18 from the first event at the British Library and 16 from the second event at Oxford University. The online questionnaire was made available in two versions, one targeted to depositors via the standard Dryad email they receive after depositing, and the other to users of the repository seeking data to reuse. This was publicised on the Dryad blog and the Jiscmail Bioinformatics listserv. While the aim was only to reach a convenience sample of early adopters rather than a statistically representative group unfortunately few responses were received; possibly as this was timed during August and early September when many academics are on vacation leave.

Questionnaire respondents identified with stakeholder groups as indicated in table 2. They were mostly representatives of publishers and journals, with relatively few authors or depositors. A broader follow-up survey would be needed to obtain statistically representative views about the factors important to authors as depositors and reusers.

The format and the criteria wording were changed following the first workshop to more accurately reflect the breakout session discussions on the 'value of Dryad' and the initial questionnaire responses, and to simplify presentation. The main changes were that:

- Two criteria ('usable deposit process' and 'ability to cite and attribute'), were not included in the second questionnaire as their relevance to the framework had already been established through gaining the highest number of 'very important' ratings in the first workshop.
- Criteria that had emerged as topics in workshop discussion were added; these were about support for a wide variety of data types, metadata standards, machine readability, repository interoperability, and peer review.
- Two policy-related issues; the clarity of curation levels, and broad community representation in governance arrangements, also emerged as discussion issues and were added to the framework without a need for further consultation.
- Two criteria relating to publishers' costs for handling supplementary material were removed as it was clear from discussion in the first workshop that cost factors would be better assessed through sustainability planning than through an assessment framework based on generic criteria and measures.
- Likert-scale questions were changed to a simpler tick box format in the second questionnaire, for all questions except those on the 'policy and process' criteria where respondents were asked how strongly they agreed or disagreed with alternative statements (see Annex 2).

Because of the questionnaire differences separate figures for each are shown in Table 2.

Responses on assessment criteria Percentage rating criteria 'very important' in workshop and online questionnaires	Publishers Journals & Societies		Authors (Depositors & Reusers)		Librarians & others		Deposit/ Reusers online
Number of responses (1)	13	6	3	7	3	3	5
4. Interaction Quality							
4.1. Deposit a wide range of data types	-	100%	-	57%	-	67%	80%
4.2. Deposit process usability	62%	-	100%		67%	-	-
4.3. Data subject to peer review	-	50%	-	14%	-	0%	20%
4.4. Discoverability	46%	50%	100%	86%	67%	67%	40%
4.5. Machine readability	-	83%	-	57%	-	67%	40%
5. Take-up and Impact							
5.1. Access stats available	62%	67%	67%	29%	33%	67%	0%
5.2. Ability to cite and attribute	62%	-	67%	-	100%	-	-
5.3. Data/ article citation impacts traceable	54%	100%	100%	86%	67%	100%	100%
5.4. Visibility/ repository interoperability	-	50%	-	14%	-	67%	40%
5.5. Evidence of community take-up	23%	33%	33%	29%	0%	33%	40%
5.6. Seen as best practice exemplar	15%	50%	67%	71%	0%	100%	60%
6. Policy and Process							
6.1. Open access licence terms apply	31%	83%	33%	71%	100%	67%	80%
6.2. Embargo period options given	-	67%	-	71	-	67%	80%
6.3. 'Trusted Digital Repository' status	46%	83%	0%	71%	67%	33%	80%
6.4. Clear curation service levels	-	-	-	-	-	-	-
6.5. Representative governance	-	-	-	-	-	-	-

Note. 1 (-) indicates criterion was not included in this questionnaire.

Table 2. Stakeholder responses on proposed assessment criteria

Table 3 below summarises sources of data appropriate for future assessment on the criteria.

Sources	(1) Quality of Interaction	(2) Take-up & Impact	(3) Policy & Process
A. Review published documentation	*	*	*
B. Stakeholder surveys	*		*
C. Usage, usability and citation analysis	*	*	

Table 3. Proposed sources of evidence

Table 4 below proposes metrics or indicators that might in future be used to assess Dryad, either to monitor its progress over time or to draw comparisons with alternative repositories.

A. Review of published documentation

An assessor reviews the documentation made publicly available by the repository and assigns a rating on the following criteria and indicators:

- Deposit a wide range of data types: Number of file formats supported for preservation; Support provided for uploading selected data to specialised repositories or databases.
- Discoverable: Relevant disciplinary metadata standards supported in repository search or navigation options; Searches may be extended to other relevant repositories; Social navigation helps users find items.
- Machine-readable: Data packages are available in an open data standard; The repository has published plans to make data packages available in an open data standard.
- Access stats available: Information is available on levels of download from and access to individual items, and to the repository's overall holdings.
- Citable and attributable: Data may be cited using a persistent ID; Guidelines are given on how to cite data items using the PID.
- Data/ article citation impacts traceable: Each data package/item is associated with a persistent identifier according to an accepted standard.
- Visible through repository interoperability: Support for repository deposit standards.
- Evidence of community take-up available: The repository makes public its designated target communities of depositors and/or reusers.
- Open access licence terms apply: data are available with minimal licence restrictions on access and reuse
- Embargo period options given: data may be deposited with an embargo on publication for a limited period subject to community norms
- 'Trusted Digital Repository' status: the repository publishes its status against a recognised Trusted Digital Repository standard.
- Clear curation service levels: the repository publishes the curation functions it performs to one or more levels of service
- Representative governance: the repository publishes its governance process

B. Stakeholder survey

Satisfaction ratings and/or comments are periodically gathered using questionnaires or interviews with representative groups of users (depositors, reusers, other stakeholders). Likert-scale agreement ratings may be useful e.g. as worded below, to monitor the perceived value of the repository on the criteria.

- Deposit a wide range of data types: I value the wide range of data types the repository supports for long-term preservation and reuse; I value the support this repository offers to upload specialized data to other repositories.
- Deposit process usable: I find it easy to deposit data in this repository; Files I need to deposit are quick to upload; I can work through the deposit steps quickly; Assistance is available and helpful.
- Data subject to peer review: The availability of data for peer review before publication benefits my research/ the research community/ research we fund/ the quality of the journal
- Discoverable: I am satisfied with the options provided to find items relevant to my needs
- Machine readable: I am satisfied with the readability of the data I can download for further analysis using software of my choice
- Seen as best practice exemplar: this repository exemplifies community best practice in data archiving
- Open access licence terms apply: I am satisfied that data is available on licence terms that maximise the potential for reuse; I am satisfied that authors have a sufficient choice of licences that may limit reuse to share-alike or non-commercial terms

- Embargo period options given: I am satisfied with the options given to depositors in this repository to embargo data for a limited period
- 'Trusted Digital Repository' status: I am satisfied this repository has plans and processes for long-term stewardship that follow accepted standards
- Clear curation service levels: I am satisfied that the repository clearly sets out what it does with data deposited to ensure it can be effectively reused.
- Representative governance: the repository is governed so that its user community has enough say in decision-making.

C. Usage, usability and citation analysis

Usage data are collected and periodically analysed; usability tests are periodically undertaken; and references to the repository and its contents are monitored

- Deposit process usable: a user can complete the deposit of a data package in less than 15 minutes; users can correctly complete deposit 95% of the time.
- Discoverable: The proportion of DOI resolution requests successfully fulfilled should be 95%; data is discoverable on the web via 3rd party indexers such as Google Scholar and ISI WoS.
- Access stats available: Annual growth in visits to landing pages; annual growth in items downloaded.
- Evidence of community take-up available: Annual growth in the number of unique and/or registered users; annual growth in data items deposited; annual annual growth in the number of depositors as a percentage of the identified target population of depositors; journals integrated with Dryad have an increasing percentage of published articles that have an associated dataset in Dryad within 6 months of publication
- Seen as best practice exemplar: favourable references to the repository as an influence on policy or practice in published articles or social media.

8.3.1 Questionnaire used in 2nd Workshop

[Note: A similar questionnaire was made available online to users of the Dryad blog, and to Dryad depositors via a link from the email inviting authors to deposit.]

Data Repositories – What Matters to Users?

Easy to deposit, to search, and to cite… What else matters when choosing an outlet such as Dryad for sharing or searching for useful research data?

Quality of the Interaction? *please tick all that are 'very important'*

- ❑ Deposit a wide range of data types, plus software or code
- ❑ Find a wide range of data types
- ❑ Search on biological metadata and keywords
- ❑ Guidance on biological metadata standards
- ❑ Machine-readable data that can be mined
- ❑ Data subject to peer review

❑ other (please describe)

Take-up and Impact? *please tick all that are 'very important'*

- ❑ Access figures shown for data deposited

- ❑ Evidence of take-up e.g. users
- ❑ Data citations are trackable
- ❑ Integrated with Databases and Institutional Repositories
- ❑ Seen as best practice exemplar in research community
- ❑ Data creators/ authors can be contacted

Policy and Processes: how far do you agree with the following? *(please tick)*

	Agree a lot	*Agree a little*	*Neither*	*Disagree a little*	*Disagree a lot*
1. Authors should use 'public domain' license terms to maximize potential for reuse (e.g. Creative Commons CC0)	❑	❑	❑	❑	❑
2. Authors should have a choice of license terms that may limit reuse (e.g. using share-alike, non-commercial license terms)	❑	❑	❑	❑	❑
3. Authors should have the option to limit access for a defined 'embargo period' e.g. a year	❑	❑	❑	❑	❑
4. There should be flexibility in the length of any embargo period e.g. for different fields	❑	❑	❑	❑	❑
5. The long-term preservation plans and processes should follow standards	❑	❑	❑	❑	❑

Finally, please identify your interest in Dryad *(please tick all that apply):*

Publisher ❑ Journal ❑ Learned Society ❑ Funding body ❑ Author who may deposit ❑ Author interested in reusing data ❑ Library or repository professional ❑ Other ❑

** Responses will inform an assessment framework being developed in the JISC funded Dryad UK Project. For queries please contact : Dr Angus Whyte, DCC, University of Edinburgh, Crichton Street, Edinburgh EH8 9LE. Please return at the end of this workshop or online by 16th Sept at* ***https://www.survey.ed.ac.uk/dryaduser***

8.3.2 Questionnaire used in 1st Workshop

☑ The Dryad Poll Please check which assessment criteria are important to you!

1. **Data use:** How important is it to assess the download statistics for data files?
   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑
2. **Workflow usability for authors:** How important is it to assess the usability for authors of the process integrating article submission and data deposit?
   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑
3. **Costs identifiable**: How important is it that publishers can identify costs of deposit and curation?

Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

4. **Cost efficiencies:** How important is it to know the extent that journals and publishers can deal with supplementary datasets more cost efficiently?

   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

5. **Acceptability of access terms:** How important is it that supplementary datasets are made available on terms agreeable to depositors?

   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

6. **Discoverability**: How important is the effectiveness of searches using non-bibliographic metadata e.g. taxa, geographic location, biological keywords?

   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

7. **Higher impact from article citation:** How important is it to monitor impact from additional article citations garnered through dataset availability?

   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

8. **Higher impact from data citation:** How important is it to monitor impact from dataset citations?

   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

9. **Curation and Preservation:** How important is it to have independent assessment of long-term stewardship planning and processes, e.g. to meet standards of 'Trusted Digital Repositories'?

   Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

10. **Best practice exemplar:** How important is it that the scientific community regard the repository as an exemplar of best practice?

    Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

11. **Journal take-up :** How important is demonstrating take-up through number of partner journals?

    Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

12. **Author take-up :** How important is demonstrating take-up through deposition rates, volumes?

    Very important ❑ Quite important ❑ Unsure ❑ Little importance ❑ Not at all important ❑

What else matters? Please add anything else you think it very important to assess:

Finally, please identify your interest in Dryad (tick all that apply):

Publisher ❑ Journal ❑ Learned Society ❑ Funding body ❑ Author who may deposit ❑ Author interested in reusing data ❑ Library or repository professional ❑ Other ❑

Thank you for your participation! Please return at the end of this workshop. Alternatively by 8th April to: Dr Angus Whyte, DCC, University of Edinburgh, Crichton Street, Edinburgh EH8 9LE