The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

Processes and Procedures for Data Publication: A Case Study in the Geosciences

Sarah Callaghan,
British Atmospheric Data Centre

Fiona Murphy,
Wiley-Blackwell

Jonathan Tedds,
University of Leicester

Rob Allan,
UK Met Office

John Kunze,
California Digital Library

Rebecca Lawrence,
Faculty of 1000 Ltd

Matthew S. Mayernik,
NCAR

Angus Whyte,
Digital Curation Centre

Abstract

The Peer REview for Publication and Accreditation of Research Data in the Earth sciences (PREPARDE) project is a JISC and NERC funded project which aims to investigate the policies and procedures required for the formal publication of research data, ranging from ingestion into a data repository, through to formal publication in a data journal. It also addresses key issues arising in the data publication paradigm, including, but not limited to, issues related to how one peer reviews a dataset, what criteria are needed for a repository to be considered objectively trustworthy, and how datasets and journal publications can be effectively cross-linked for the benefit of the wider research community. PREPARDE brings together a wide range of experts in the research, academic publishing and data management fields both within the Earth Sciences and in the broader life sciences with the aim of producing general guidelines applicable to a wide range of scientific disciplines and data publication types. This paper provides details of the work done in the first half of the project; the project itself will be completed in June 2013.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v8i1.253

Introduction

Data has always been the foundation of scientific progress, although up until recently it has been difficult to share and scrutinise. The Internet has provided the opportunity for large (and small) datasets to be passed around quickly and easily, and made available for anyone to use. These possibilities, though exciting, are also unnerving to researchers who spend a great deal of time and effort creating and managing datasets. Once a dataset is made freely downloadable from a webpage, it is hard to track how the data is being used and by whom, and limited academic credit is currently provided to the dataset creators. For researchers who want to use the data, there are no guarantees that the data that they find on a webpage is in a standard (or even understandable) format, or hasn't been changed since the last time they downloaded it.

Data citation is proposed as a mechanism for providing the dataset creators with the recognition they deserve for creating and managing the data, through the development of citation metrics (currently under development by Thomson-Reuters and others). Data publication can then piggy-back on the existing journal infrastructure to provide a "stamp of approval" in the form of scientific peer review (Figure 1). Formalised citation standards (such as using DOIs) also provide mechanisms to encourage version control of datasets.

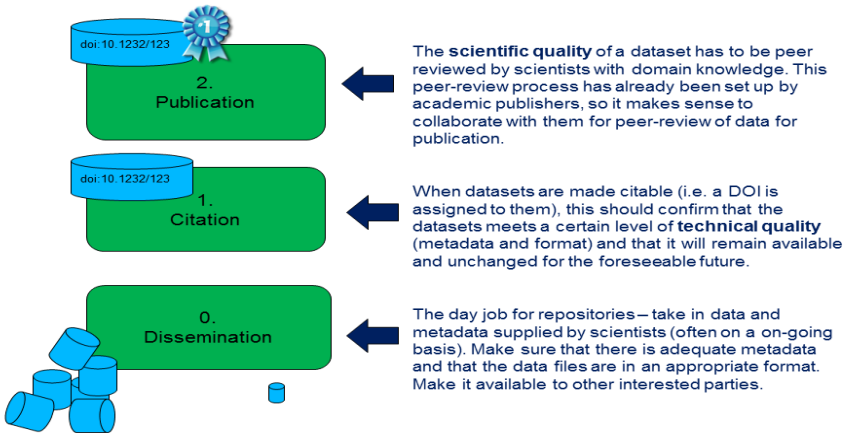

Figure 1. Dissemination, citation, and publication of data correspond to three levels of "quality". Note that here, publication signifies a formal process of making something public after adding value for the consumer, e.g., after peer review, as distinct from informal dissemination, which encompasses any process that simply makes data available for download.

Formal publication of data provides a service over and above the simple act of posting a dataset on a website, in that it includes a series of checks on the dataset of either a technical (format, metadata) or a more scientific (is the data scientifically meaningful?) nature. Formal data publication also provides the data user with certain assurances about data persistence, and provides a forum for the dataset to be found and evaluated – an essential part of the scientific process. The technological infrastructure needed for data publication exists and is mature in certain areas of science, particularly in the Earth and Geo-sciences. What is now required are the

processes and procedures to ensure the smooth running of data journals, and the means to encourage authors to submit their data for publication.

Publishers are increasingly expressing an interest in publishing data, and of particular relevance to the PREPARDE (Peer Review for Publication and Accreditation of Research Data in the Earth sciences) project is that 2012 has seen the launch of the Geoscience Data Journal (GDJ) from Wiley which:

> "...provides an open access platform where scientific data can be formally published, in a way that includes scientific peer review… An online-only journal, GDJ publishes short data papers cross-linked to, and citing, datasets that have been deposited in approved data centres and awarded DOIs."[1]

A "data paper" describes a dataset, giving details of its collection, processing, calibration, software, file formats etc., without the requirement of novel analyses or ground breaking conclusions, allowing the data paper to be published rapidly after the completion of the dataset. This encourages other users either to cite the data directly (as publication requires the dataset to have a DOI or other permanent identifier), or to cite the data using the data paper as a proxy. Additionally, the data paper allows the reader to understand when, how and why the data were collected, and the research context in which the dataset was generated.

The launch of GDJ provides a rare opportunity to build the new processes and procedures required for data publication into a new data journal title. In order to take advantage of this, the PREPARDE project has been funded by JISC and NERC to investigate issues such as data paper submission workflows, cross-linking between articles and datasets, data repository accreditation and the scientific peer review of data.

PREPARDE is investigating the key organisational, procedural and economic challenges for data publication. For example:

- What journal and repository policies are required to achieve greater levels of data sharing, citation and linkages between publications and datasets?
- What partnerships between journals, data centres and research organisations are necessary to establish sustainable data publication solutions, and what business models are appropriate to sustain them in the long term?
- What characterises a suitable, trustworthy repository?
- What peer review of data (technical and scientific) is appropriate, and at which stage(s) of the publication process, to include acceptable levels of validation and error estimation?

PREPARDE brings together academic researchers, journal publishers and data centres to address these issues and to produce guidelines and project outputs that aim to be applicable across a wide range of research data publication.

[1] Geoscience Data Journal: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-GDJ3.html

 doi:10.2218/ijdc.v8i1.253

The PREPARDE project team includes a wide range of partners including academic institutions, learned societies, data centres and commercial publishers, both nationally and internationally. It aims to develop the mechanisms required to enable data to be identified, cited and published with confidence. This involves investigating barriers and drivers to data publishing and sharing, peer review, and re-use of geoscientific datasets. The project is also looking at all these issues in relation to a broader set of fields beyond geosciences, including a collaboration with *F1000Research* which publishes in the life sciences.

A key goal of the PREPARDE project is to ensure that it reaches out to, and is informed by, other related initiatives on a global basis – in particular, those interested in developing long-term sustainable policies, processes, incentives and business models for managing and publishing research data. Interaction with data producers is vital in order to ensure that datasets are prepared appropriately for publication. Similarly, community engagement and buy-in to the data review processes are essential if scientific review of data is to become the norm, rather than the exception.

The main benefits anticipated to arise from PREPARDE are:

- For researchers, it will provide an appropriate forum to make their data reusable and sharable, and to receive credit for doing so while allowing them to directly influence data publication guidelines.
- For publishers, the project will enable development of data publication guidelines that have more detail, are better implemented, have greater community input than currently exist and are extensible to a wide range of journals.
- For repository partners, collation of best practices for repository accreditation will enable them to improve their data management processes, and cross-linking with journals will improve their visibility and prestige.
- For the wider community of stakeholders (including funders, institutions, learned societies, businesses and the public) it will build support for the full range of scholarly communication and dissemination.

Workflows and Cross-Linking Between Data Centres and Data Journals

Workflows

The Geoscience Data Journal has built close relationships with well-known and well-respected data centres in the Earth Sciences. Part of the PREPARDE project involves studying workflows for data centre ingest (in these examples, the British Atmospheric Data Centre – BADC – and the National Centre for Atmospheric Research – NCAR) and taking the dataset from submission to publication. By investigating these workflows, even though they take place at separate institutions, it is possible to identify points where effective cross-linking can enhance the operation of both partners. As a control, the workflows from a non-data publisher (International Journal of Digital Curation) were also scrutinised. Work on this topic is on-going, and

will be the subject of a workshop to be held in 2013. The following describes early impressions from the workflows.

Data repository workflows may vary according to the type of data submitter. For example:

- "Engaged submitter" – dataset author is engaged in the process of dataset ingestion into the archive and will answer questions and provide metadata and supporting documentation (see Figure 2 for an example workflow). Datasets from engaged submitters are most likely to be assigned with DOIs after the ingestion process is completed.

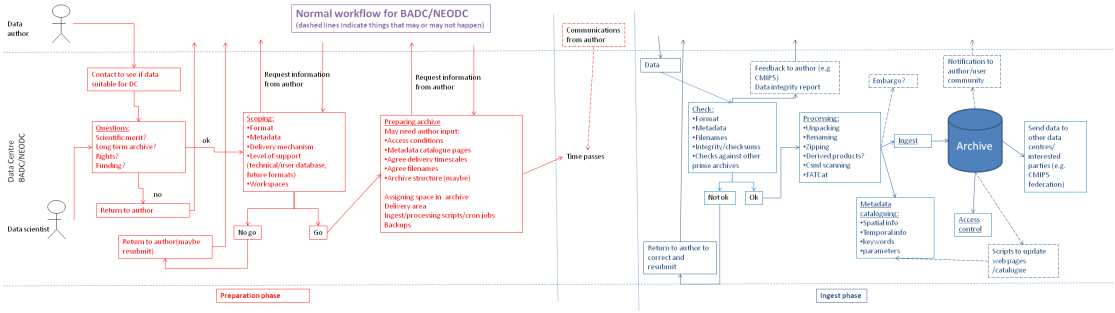

Figure 2. Example BADC data ingestion workflow (for the "engaged submitter").

- "Data dumper" – the dataset is provided to the data centre "as-is" with no further supporting information, metadata or contact with the author. In some cases, this is legacy data where the data centre is archiving it to save it from deletion. These datasets are unlikely to be awarded DOIs, as they probably do not meet the technical requirements for DOIs. However, if it is determined that these datasets are scientifically important, then some effort may be spent on digging up more metadata and/or cleaning up the dataset, and they then might be awarded a DOI.
- "Third party data request" – at the BADC this is when a researcher asks the data centre to broker a transfer of data between them and a third party (e.g. the UK Met Office), where the data centre coordinates the transfer of data from one party to another via the data centre's already established channels. DOIs may or may not be assigned to these datasets, depending on the licensing conditions associated with the transfer of the data between the researcher and the third party, as well as the conditions of storage of the data in the data centre.
- At the BADC (and all the NERC data centres) DOIs get assigned at the end of the ingestion process, and after a few more checks of the dataset to ensure it meets the technical quality requirements to be assigned a DOI. However, F1000R assign DOIs and then employ a post publication peer review model that we will compare.

 doi:10.2218/ijdc.v8i1.253

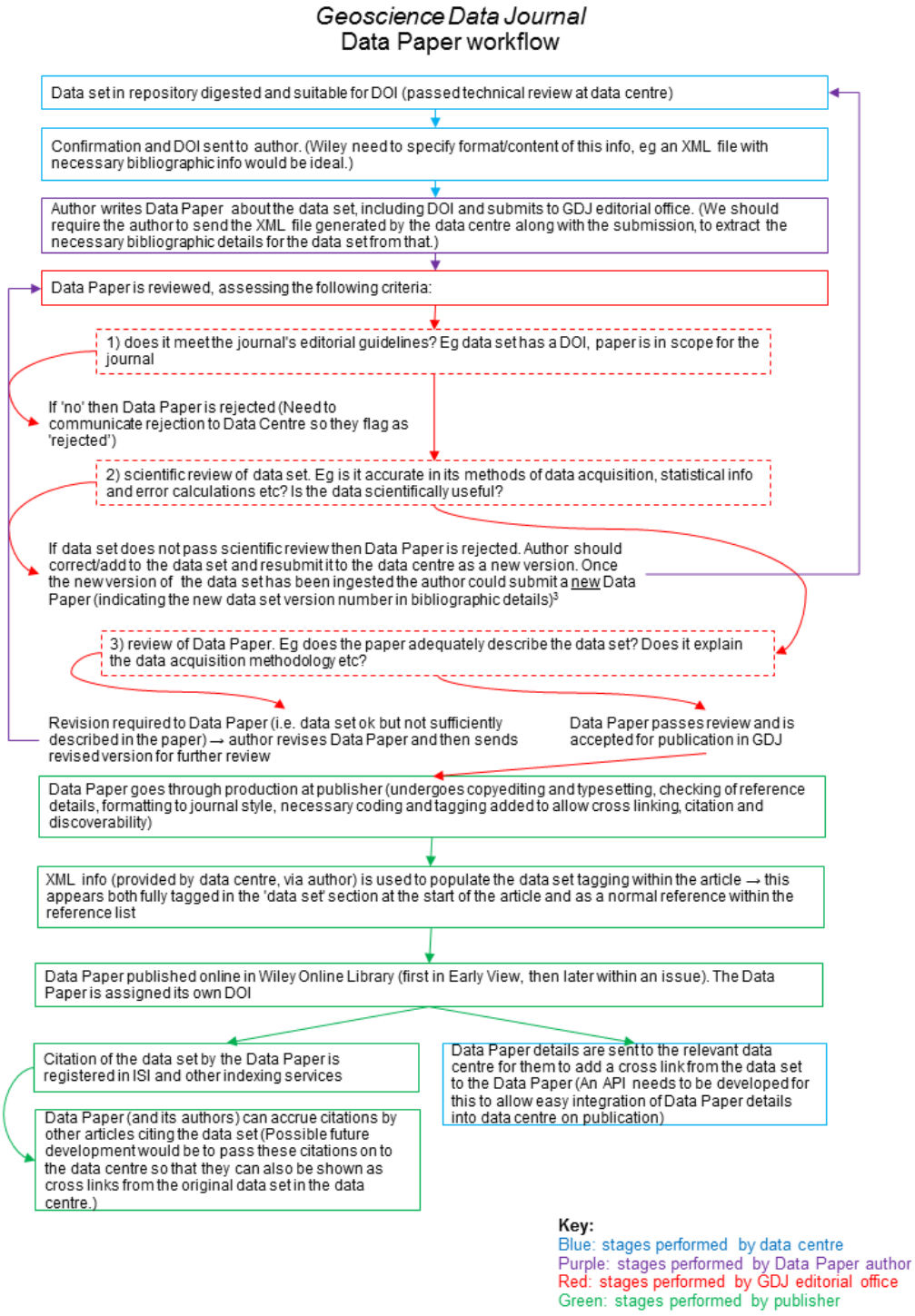

Figure 3. Example GDJ data publication workflow.

- Cross-linking between BADC and GDJ, as shown in Figure 3 in the box "XML info (provided by data centre, via author)…", is used to populate the data set tagging within the paper. Hence, it is possible to engineer a link and provide the XML information without the input of the author by simply using the DOI to connect to the dataset metadata record and ingesting the appropriate metadata for the data paper that way. This means that the metadata collected by the data

centre needs to be easily mapped to the metadata needed by the journal. An intermediate step would involve automatically ingesting the appropriate metadata from the data centre into a webform that the data paper author could edit/add to, so the resulting metadata could then be shared between the data centre and the journal. This would mean that the dataset author would not have to provide information to the journal that had already been provided to the data centre, and would be able to improve/review the metadata held by the data centre at the same time as the data paper metadata is checked.

- The data journal has several places where it needs to communicate back to the data centre the results of reviews, such as when the data paper is published, citations of the data paper etc, so that further cross-links can be made to enrich the dataset. It is important not only that the data paper cite the relevant datasets, but also that the data centre links back to data papers that result from those datasets.
- Results of journal processes may impact the dataset that is stored in a data centre. For example, a data review might require a correction that in turn requires that a new version of a dataset (with a new DOI) be created and stored.[2] It might be the case that a dataset needs to be withdrawn from the data centre, with a redaction notice added to the landing page for the DOI. In this event, the data centre would need to notify the data journal that the data is no longer available, but it would be up to the journal to update the data paper and/or paper metadata to reflect this. Likewise, it may be that, following peer review of the data by the journal, the dataset should be withdrawn from the data centre and so again, close communication between the journal and the data centre are required throughout the review processes by both parties, to ensure both are kept closely up to date and any links between the two are appropriately maintained over time.

Cross-Linking

Figure 4 shows a mockup of a data paper as it is likely to appear on the GDJ site. What is important to note is that the link to the dataset (via a DOI) is provided right at the very top of the paper, before even the abstract, highlighting the dataset's importance to the paper. The dataset citation is also included in the reference list at the end of the paper for two reasons: firstly to take advantage of citation counting systems that use the reference list as their source of information, and secondly to affirm the importance of the dataset as a first class research output.

This link at the top of the paper is the foundation of cross-linking between the data repository and the data paper (although other ways are possible), and can enrich the paper and user experience through interactive services that providers build on top. For example, for a dataset spanning a geographic region, location coordinates can be used to plot the dataset on a map. An example of this can be seen in Elsevier's Article of the Future.[3] Note that the PREPARDE project will mainly be concentrating on the

[2] This may depend on the severity of the change to the dataset. Note that there is some policy disparity between DOI assigners (data publishers, data centres, aggregators, etc.) on what thresholds of change to a dataset should trigger the assignment of a new DOI.

[3] An example of Elsevier's Article of the Future can be viewed at http://www.articleofthefuture.com/S0031018208004690/

policies and procedures for data publication, so examples of cross-linking beyond simple citation via DOI are out of scope for this project. Further information on citing and linking research data can be found in Ball and Duke (2012).

Figure 4. Sample layout of a data paper on the GDJ site

Data Centre Accreditation

Data centre accreditation is an important factor for data publication, especially in the case of the GDJ, where the data referred to in the paper is stored in a data repository external to the journal systems. This raises issues of trust, in that the links between the data paper and the dataset have to be maintained for the long term, and the data need to be archived and curated in a manner that safeguards the scientific record.

Repositories should:

- Have long term data preservation plans in place for their archive;
- Actively manage and curate the data in their archive;
- Provide landing pages giving extra information about the dataset (metadata) and information on how to access the data;
- Use persistent, actionable links (e.g., DOIs, ARKs) to cite data held in their archive;
- Resolve cited dataset links to landing pages.

There are several data centre accreditation schemes proposed in the literature, of varying levels of complexity. One is TrustedDigitalRepository.eu, which is a collaboration between the Data Seal of Approval, the Repository Audit and Certification Working Group of the CCSDS, and the DIN Working Group "Trustworthy Archives – Certification." They propose a three-tiered framework[4] which will consist of a sequence of three levels of increasing trustworthiness:

- Basic Certification is granted to repositories which obtain DSA[5] certification;
- Extended Certification is granted to Basic Certification repositories which in addition perform a structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;
- Formal Certification is granted to repositories which, in addition to Basic Certification, obtain full external audit and certification based on ISO 16363 or equivalent DIN 31644.

For a data journal editor, the main question that needs to be answered is: "Is this repository trustworthy?" These accreditation schemes go into considerable detail about many aspects of a repository's activities. Does the journal editor need to know the details of what is examined in each certification scheme, or only a broad outline of what is audited and whether the requirements are appropriate for asserting the trustworthiness of the resources being described in a data paper? An additional complication is that the requirements for trustworthiness may be beyond the expertise of the journal editors, and may vary from journal to journal, or from subject to subject.

Can learned societies and other organisations intermediate effectively by mapping community standards, (e.g., for context and provenance information) to the data review needs of both journal editors and repository accreditation standards? Since public research funding bodies increasingly expect institutions to be accountable and potentially responsible for data management, can they play an effective role in assuring the quality of data submitted to repositories and journals? The topic of repository accreditation was addressed in more detail at a workshop following the International Digital Curation conference in January 2013, and a workshop report will follow.

[4] TrustedDigitalRepository.eu framework: http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

[5] Data Seal of Approval: http://www.datasealofapproval.org/

 doi:10.2218/ijdc.v8i1.253

Scientific Peer Review of Data

Scientific peer review of data is acknowledged to be one of the more challenging aspects of data publication. In 2010, the *Journal of Neuroscience* decided to stop allowing any supplementary files for their articles because they felt their referees were getting overwhelmed with the volume of data and other files that required refereeing in addition to the main article.[6] Peer review of the main text and figures in papers is already a time-consuming and difficult task, but it is well established and well understood. However, the notion of peer review of data can be confusing and costly.[7] Exactly what is reviewed? The raw data itself? The metadata? The data paper? All of them? What is practical to ask a referee to do? Should this occur pre- or post-publication even?

Lawrence et al. (2011) provide a few working examples of peer review of data, along with a data review checklist, asking questions to which the data reviewer can give definitive answers. Such checklists provide a structured way of producing a data peer review process, and guide a new reviewer to the things that should be considered as part of the review. However, checklists are often generic, and may have to be tailored specifically for the subject domain.

It may be helpful to split peer review up into separate phases carried out by different people. In the case where a dataset is held in a trusted repository, technical checks are carried out when the data is ingested into the data centre. Such checks usually include ensuring that the data is in an appropriate and standard format, with full metadata. This simplifies the work of the scientific reviewer who knows that they'll already be able to find the dataset, open it, and perhaps even use tools and viewers provided by the data centre to investigate the data. The scientific reviewer can then concentrate on judging whether the data and metadata provided are scientifically meaningful, without having to waste time finding the right software to open a file.

On the issue of post- versus pre-publication peer review, in order to be able to provide an informed opinion on how good a dataset really is usually requires the data to be interrogated, reused, and maybe even replicated. It is impractical to expect a busy referee to do this in the tight timeframe typically required prior to publication. However, this is exactly what other research groups working directly in the field are likely to want to do with the data following publication, and so it makes sense to capture this feedback for the benefit of others. Hence, some of the most effective and complete peer review of a dataset is likely to occur after publication, from other researchers either trying to replicate the data or reusing it in further studies.

In the life sciences field, the new Open Access journal *F1000Research* is using post-publication peer review only, in which the article and data undergoes a quick internal check and is then published with a DOI, together with all the data underlying the results (this is mandatory). The article and data then goes immediately into a formal, invited peer review process (completely open and signed), but in addition, any researcher who states their full name and affiliation can also comment on the article and data. This approach means the process of review never ends, and hence at any

[6] The full announcement can be found at: http://www.jneurosci.org/content/30/32/10599.full

[7] Not to be undertaken lightly; a partner approach that trades formal peer review for data source review to achieve rapid data publication may be found in Kunze et al. (2011).

point when someone tries to reuse or play with the data, they can add their detailed feedback on the data, better mirroring how scientific research is actually conducted. The topic of how to peer review data will also be the subject of a specific individual workshop, to be held in March 2013.

Concluding Remarks and Future Work

The PREPARDE project is ambitious in aiming to bring together a wide range of stakeholders to produce a set of policies and procedures required for the operation of a data journal. To limit the scope of the project, the main focus is on the Earth Sciences, specifically the Geoscience Data Journal, but through collaborations we will examine these issues in the wider research context, and hence the conclusions drawn and guidelines created from this project will be generalizable to other areas of research.

As always, community engagement in this work is essential for the guidelines to meet the needs of all users. To that end, the project team welcome all comments and interactions. The PREPARDE project website can be found at http://proj.badc.rl.ac.uk/preparde.

Acknowledgements

The project is led by the University of Leicester. The support of JISC and NERC in funding the PREPARDE project is gratefully acknowledged.

References

Lawrence, B., Jones, C., Matthews, C., Pepler, S. & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation 6*(2), 4-37. doi:10.2218/ijdc.v6i2.205

Ball, A. & Duke, M. (2012). *Data Citation and Linking.* DCC Briefing Papers. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/

Kunze, J., Cruse, P., Hu, R., Abrams, S., et al. (2011). *Practices, trends, and recommendations in technical appendix usage for selected data-intensive disciplines.* California Digital Library. Retrieved from http://n2t.net/ark:/13030/c7jw86m55