JISC

PREPARDE
4.4: Produce a business model addressing sustainability of the partnerships developed in this work package and funding for future developments in cross-linking.

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences		
Project Hashtag	#preparde		
Start Date	1 July 2012	**End Date**	30 June 2013
Lead Institution	University of Leicester		
Project Director	Dr Jonathan Tedds		
Project Manager	Dr Sarah Callaghan		
Contact email	sarah.callaghan@stfc.ac.uk		
Partner Institutions	University of Leicester British Atmospheric Data Centre (BADC) US National Centre for Atmospheric Research (NCAR) California Digital Library (CDL) Digital Curation Centre (DCC) University of Reading Wiley-Blackwell Faculty of 1000 Ltd		
Project Webpage URL	http://proj.badc.rl.ac.uk/preparde/wiki		
Programme Name	*Managing Research Data*		
Programme Manager	Simon Hodson		

Document Information			
Author(s)	Fiona Murphy		
Project Role(s)	Project Participant		
Date		**Filename**	
URL	*If this report is on your project web site*		

Access	☐ Project and JISC internal	☒ General dissemination

Document History		
Version	**Date**	**Comments**
1	15 January 2014	First draft
6	4 July 2014	Final draft

Currently, it is very much the exception rather than the norm to publish primary research material cross-linked with relevant datasets. However, given recent developments in funder and policy-maker emphasis, we believe it will be increasingly important (indeed critical), for researchers to include such links as part of the process of publication.

Potential Positive and Negative Consequences of Cross-linking

The various impacts of cross-linking generally follow those of data publication and data management, namely:

Positive

1. Supports the potential to re-use research data by improving discoverability, permanent availability (subject to licensing), metadata, improving transparency and inter-disciplinary usage.
2. Provides academic credit to the data gatherers.
3. Progresses the 'Open Science' agenda which is increasingly gathering pace amongst researchers and policy-makers at a global level.[1]

Problematic

1. Increasingly complex workflow issues – each repository could have a different workflow
2. Hits the 'sharing' barrier – currently there are insufficient incentives to share embedded in the system
3. Lack of standards with respect to citation behaviours (what to cite and when), quality control, metadata, persistence of the repository's funding and mission.

Notwithstanding these issues, which are addressed elsewhere by PREPARDE, the development of a cross-linking culture in scholarly publication is a key priority, given its potential benefits for research impact.

---

[1] See for instance 'Science as an Open Enterprise' Report published by The Royal Society, June 2012.

A viable cross-linking ecosystem depends on the cooperation of all the key partners – data centres, publishers, authors, journals and funders – within the workflow. As well as being aware of each other's roles, motivations and requirements, there also needs to be:

- A common understanding of the value-add cross-linking affords
- Rewards and barriers to future funds and career advancement dependent on compliance or otherwise
- The development of metrics upon which real decisions are made
- A common understanding of when authors should cite datasets directly, data papers and/or primary research articles
- If it can be shown objectively, research evidence to support the additional effort involved (i.e. showing increased impact)
- Widespread engagement with research communities
- Opportunities to embed information about both dataset and publication seamlessly into publishing workflows
- Workflows need to allow for post-, as well as simultaneous-to-publication, linking
- A central registry managing the bilateral links between datasets and publications
- An interim measure of standard Memoranda of Understanding between data journals (publishers) and repositories. This MoU would relate to the accreditation guidelines or 'agreed standards for approved repositories' mentioned under 'sustainable business model'.

Sustainable Business Model

1. A comprehensive, persuasive roadmap showing clear areas of responsibility in terms of workflow and potential financial outlay should be compiled (see Appendix for sample draft).
2. Agreed standards of quality control and metadata collection.
3. Agreed practices for dataset/datapaper/primary research citation, accreditation and metrics collection
4. Agreed standards for 'approved repositories'
5. Agreed standard workflows for repositories and data publishers

Proposed Actions

- The PREPARDE project team to continue engaging with thought leaders within research communities, including learned societies, key research groups, high profile academics involved with policy making.
  - Preliminary list of specific entities includes: American Geophysical Union, European Geophysical Union, Integrated Earth Data Applications, JISC, RCUK, Research Data Alliance, Thomson Reuters[2], National Science

---

[2] See recent report: 'Unlocking the Value of Research Data' published by Thomson Reuters Industry Forum.
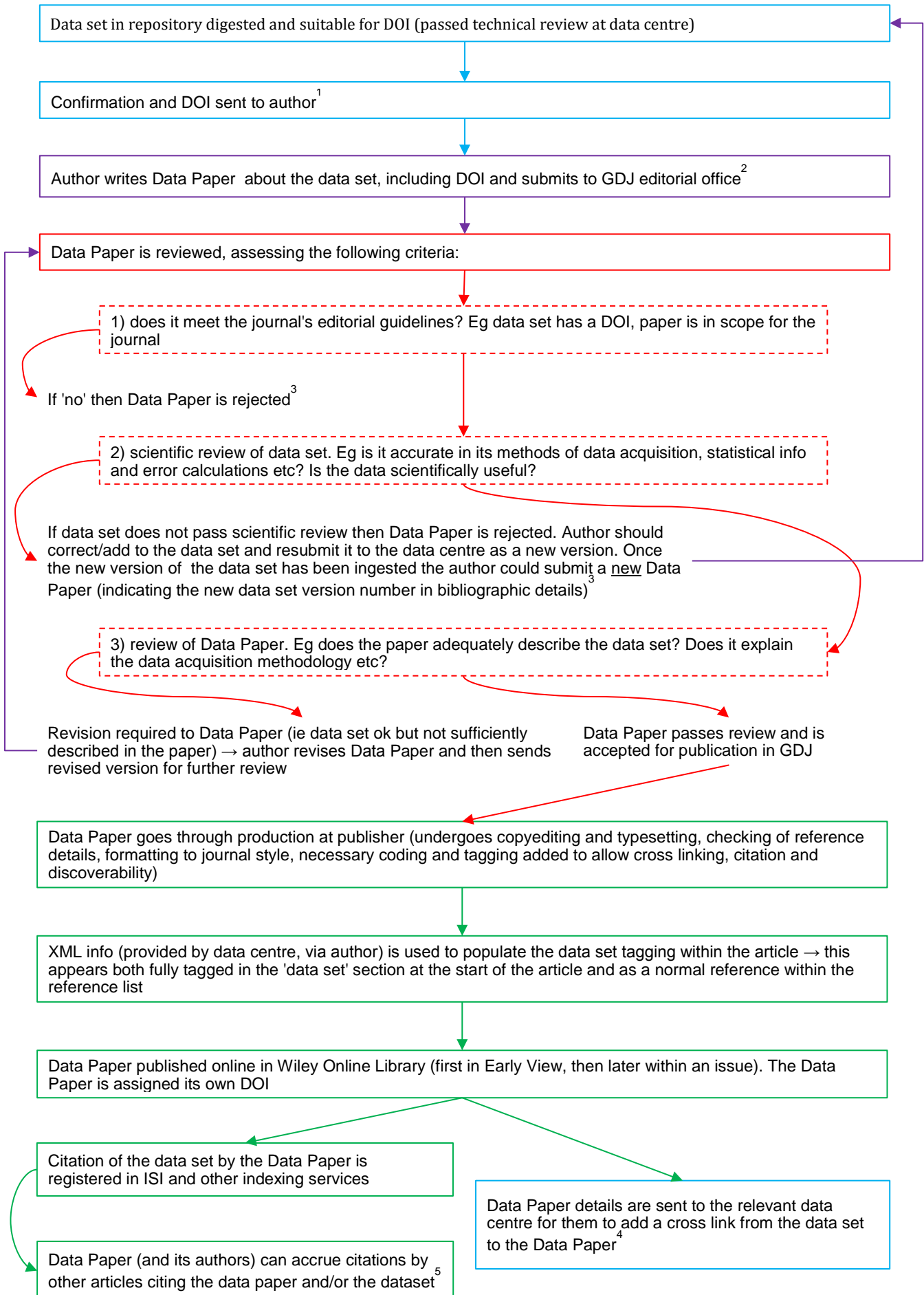
Foundation, EarthCube, CODATA, WDS, EU (via COST, CNECT, OpenAIRE), Association of Learned and Professional Society Publishers, amongst others.

- The team is active within Research Data Alliance, CODATA and the World Data System Working, Task and Interest Groups that are producing recommendations for global action on this and other data publication issues.
- Working with other stakeholders, the team to take active steps towards evolving and disseminating best practice guidance on data publication, cross-linking and citation issues. For example, we have developed the high-level standard "Cite what you use"[3], together with further background clarification.
- Take every opportunity to work with funders (e.g. SIM4RDM, EU CNECT open research data consultation, etc.) to contribute to the case for supporting:
  - direct action in the form of resources to contribute to the construction of such a registry
  - indirect action such as putting appropriate mandates in order to incentivise researchers to properly manage and share research data
  - the need for funds to be made available to build expertise via training and to enable support services within funding or institutional infrastructures. Once built, it is expected that minimal or no funding will be required to maintain the system.
- It is critical to engage with CrossRef regarding the registry. As a non-profit, independent organisation working in an extremely similar space this is the obvious entity to take the initiative forward. Initial contact has been made. Further actions may be required to progress this.

---

[3] Blogpost by Sarah Callaghan: http://citingbytes.blogspot.co.uk/2014/01/cite-what-you-use.html

Appendix: Sample workflow including cross-linking and financial considerations
Geoscience Data Journal

Data set in repository digested and suitable for DOI (passed technical review at data centre)

Confirmation and DOI sent to author[1]

Author writes Data Paper about the data set, including DOI and submits to GDJ editorial office[2]

Data Paper is reviewed, assessing the following criteria:

1) does it meet the journal's editorial guidelines? Eg data set has a DOI, paper is in scope for the journal

If 'no' then Data Paper is rejected[3]

2) scientific review of data set. Eg is it accurate in its methods of data acquisition, statistical info and error calculations etc? Is the data scientifically useful?

If data set does not pass scientific review then Data Paper is rejected. Author should correct/add to the data set and resubmit it to the data centre as a new version. Once the new version of the data set has been ingested the author could submit a new Data Paper (indicating the new data set version number in bibliographic details)[3]

3) review of Data Paper. Eg does the paper adequately describe the data set? Does it explain the data acquisition methodology etc?

Revision required to Data Paper (ie data set ok but not sufficiently described in the paper) → author revises Data Paper and then sends revised version for further review

Data Paper passes review and is accepted for publication in GDJ

Data Paper goes through production at publisher (undergoes copyediting and typesetting, checking of reference details, formatting to journal style, necessary coding and tagging added to allow cross linking, citation and discoverability)

XML info (provided by data centre, via author) is used to populate the data set tagging within the article → this appears both fully tagged in the 'data set' section at the start of the article and as a normal reference within the reference list

Data Paper published online in Wiley Online Library (first in Early View, then later within an issue). The Data Paper is assigned its own DOI

Citation of the data set by the Data Paper is registered in ISI and other indexing services

Data Paper details are sent to the relevant data centre for them to add a cross link from the data set to the Data Paper[4]

Data Paper (and its authors) can accrue citations by other articles citing the data paper and/or the dataset[5]

Notes

[1] Publisher – in this instance Wiley – needs to specify format/content of this info, e.g. an XML file with necessary bibliographic info would be ideal.

[2] Publisher should require the author to send the XML file generated by the data centre along with the submission, to extract the necessary bibliographic details for the data set from that.

[3] Need to communicate rejection to Data Centre so they flag as 'rejected'.

[4] It could be argued that this is currently a joint responsibility of both the Publisher and the data centre – which is problematic for several reasons. Firstly, an API needs to be developed for this to allow easy integration of Data Paper details into data centre on publication. Secondly, the bilateral and fragile connection between the two entities needs to be strengthened and scaled up.

[5] Note that the authorlist of the data paper(s) and the corresponding dataset(s) may or may not be identical. A future development would be to feed these citations between data centres and publishers via fully functioning bi-directional cross links. Ultimately the Registry would become involved with monitoring the cross-linking mechanisms as well as harvesting metrics.

Key to workflow diagram:

Blue: stages performed by data centre – could be maintained by funder, directly or indirect via grant slicing. The funder would then be in a good position to assess the content and impact of such datasets. Would imply that project id should be included as essential metadata for both paper and dataset. Regarding the final box in the flowchart above, publishers, data centres and funders need to develop a close understanding of how each stakeholder can potentially contribute to and benefit from a Registry. A jointly funded, properly governanced system with a mixture of private and public funding would likely prove the most effective solution.

Purple: stages performed by Data Paper author. Given that data papers might be written by a different set of project team members from primary research papers, financial support and/or training may need to be provided, as well as the reward/incentive system re-examined. Such support might be provided by the funder or the institution. The publisher may be able to provide some of the training. This responsibility might also be taken up in part by Learned Societies.

Red: stages performed by GDJ editorial office. To date, the journal's own business model (generally subscriptions, occasionally a learned society's own funds) has been used to support the peer review and editorial processes. Given that peer review itself is changing, however, this may no longer be a given. New models and commercial players are emerging so that peer review and a particular publication are no longer inextricably linked. Some groups of journals are beginning to share peer review responsibility, new companies are offering peer review not linked with specific publication outlets. Likely new tools will come into the market (open source?) to enable visualisations, and other methods of assessing and verifying data.

Green: stages performed by publisher. The production and publication processes are key publisher functions. At present there does not seem to be a need to test whether this should continue to be the case.