



Perspectives on the Role of Trustworthy Repository Standards in Data Journal Publication

IASSIST Cologne, 31 May 2013

Angus Whyte, Sarah Callaghan,
Jonathan Tedds, Matthew S. Mayernik,
and the PREPARDE project team

#preparde

a.whyte@ed.ac.uk



Aims

1. Introduce the PREPARDE project
 - Data journal and repository links
 - Data peer-review
 - Repository trust accreditation *
2. Repository certification background
 - Why relevant to data journals
 - Standards developed
 - Issues being discussed
1. PREPARDE Guidelines
 - Input to them from IDCC workshop
 - Hopefully also your comments...

Q. What should repositories, depositors and journals expect from one another?

PREPARDE Guidelines

Q. What are use cases for data journals in social sciences?

Q. What support should institutions offer?

PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences

Lead Institution: University of Leicester

Partners

- 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

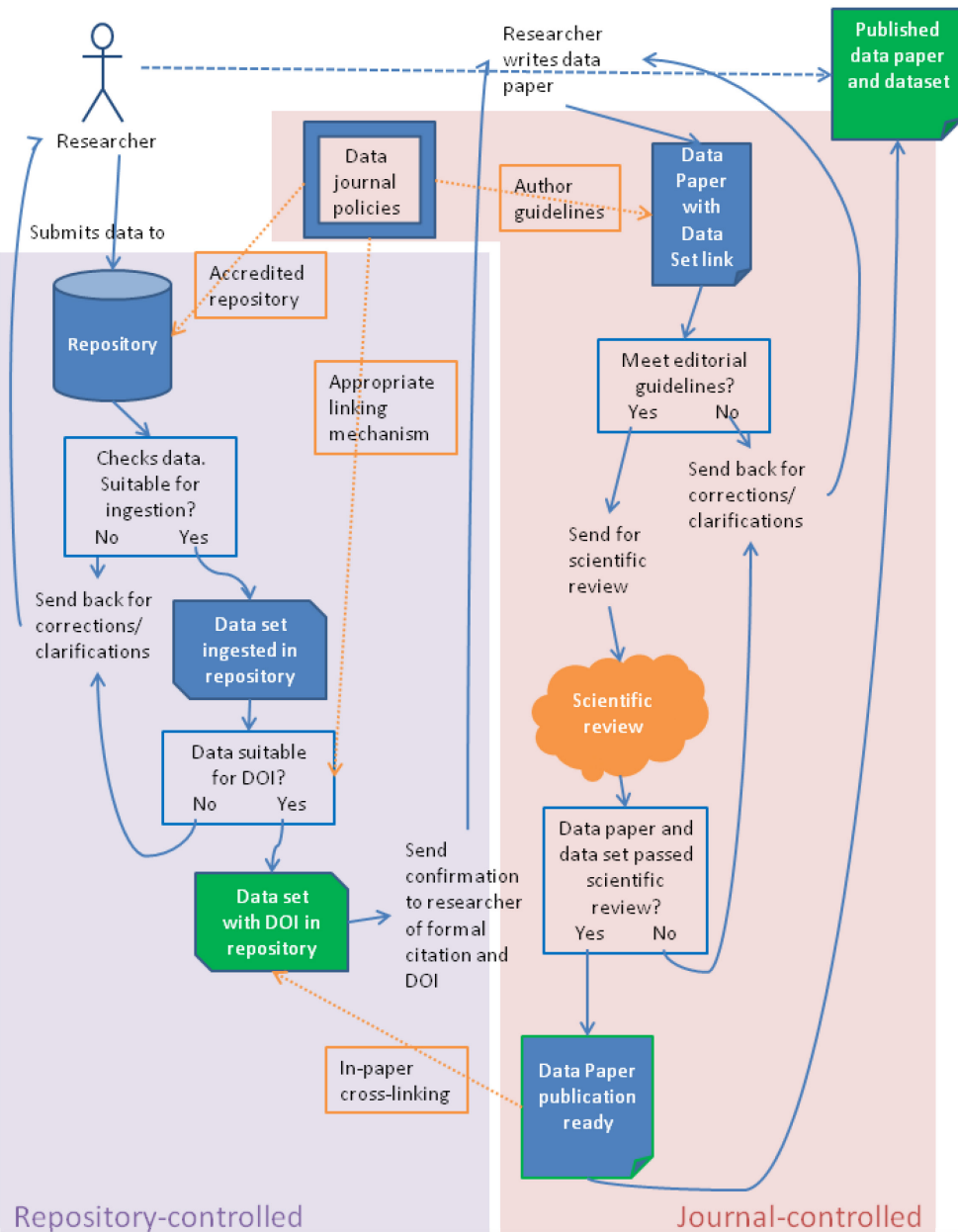
Project Lead: Dr Jonathan Tedds (University of Leicester, jat26@le.ac.uk)

Project Manager: Dr Sarah Callaghan (BADC, sarah.callaghan@stfc.ac.uk)

Length of Project: 12 months

Project Start Date: 1st July 2012

Project End Date: 31st June 2013



PREPARDE topics

3 main areas of interest (in orange)

1. Workflows and cross-linking between journal and repository
2. Repository accreditation
3. Scientific peer-review of data

Main aim: to put in place the policies and procedures needed for data publication in the Geoscience Data Journal and to generalise those policies for application outside the Earth Sciences.

Why: Reasons for citing and publishing data



<http://www.evidencebased-management.com/blog/2011/11/04/new-evidence-on-big-bonuses/>

- **Pressure** from (UK) **government** to make data from publicly funded research available for free.
 - Scientists want attribution and credit for their work
 - Public want to know what the scientists are doing
- Research **funders** want reassurance that they're getting **value for money**
 - Relies on peer-review of science publications (well established) and data (not done yet!)
- Allows the wider **research community** to **find and use** datasets, and understand the **quality** of the data
- Extra **incentive** for scientists to submit their data to data centres in appropriate formats and with full metadata

How: *Geoscience Data Journal*, Wiley-Blackwell and the Royal Meteorological Society

- Partnership to develop a mechanism for the formal publication of data in the Open Access *Geoscience Data Journal*
- GDJ publishes short data articles cross-linked to and **citing datasets that have been deposited in approved data centres** and awarded DOIs or other permanent identifier.
- A **data article describes a dataset** collection, processing, software, file formats, etc., without the requirement of novel analyses or ground breaking conclusions.
 - the **when, how and why** data was collected and what the data-product is.



Author Guidelines

Dataset submission

“authors must complete the following two-tiered process:

The dataset, along with supporting metadata, must be formally archived in a *Geoscience Data Journal* approved repository or data centre (and preferably have been assigned a [digital object identifier \(DOI\)](#))

...An approved repository is one that is commonly used by the scientific community it supports, has a formal data management policy in place, and can mint a DOI or provide a stable URL and unique identifier for the dataset. “



Current approved repositories are:

[3TU.Datacentrum](#)

[British Atmospheric Data Centre \(BADC\)](#)

[British Oceanographic Data Centre \(BODC\)](#)

[CSIRO Data Access Portal](#)

[Environmental Information Data Centre \(EIDC\)](#)

[Figshare](#)

[National Geoscience Data Centre \(NGDC\)](#)

[NERC Earth Observation Data Centre \(NEODC\)](#)

[PANGAEA](#)

[Polar Data Centre \(PDC\)](#)

Author Guidelines

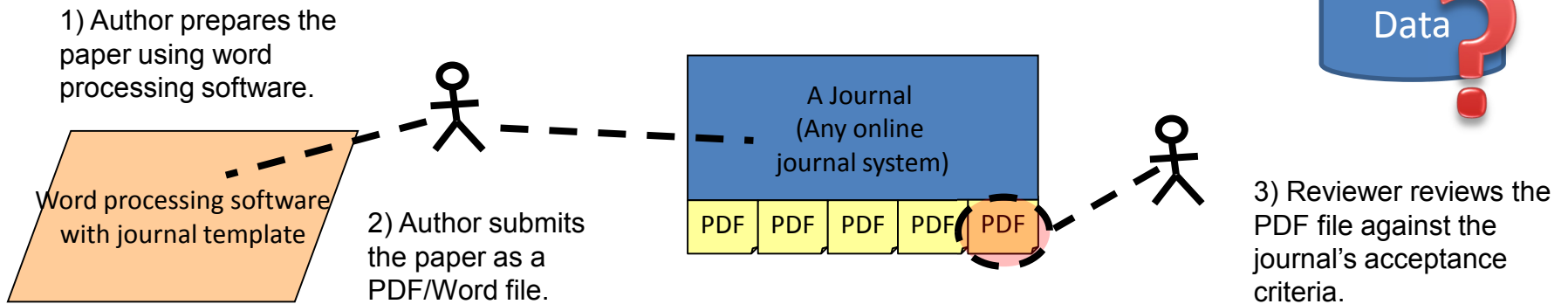
Dataset submission

... Subject to satisfactory reviews of both dataset and paper, *Geoscience Data Journal* will publish the data description paper, along with a link to the underlying dataset (usually by means of the dataset's DOI).

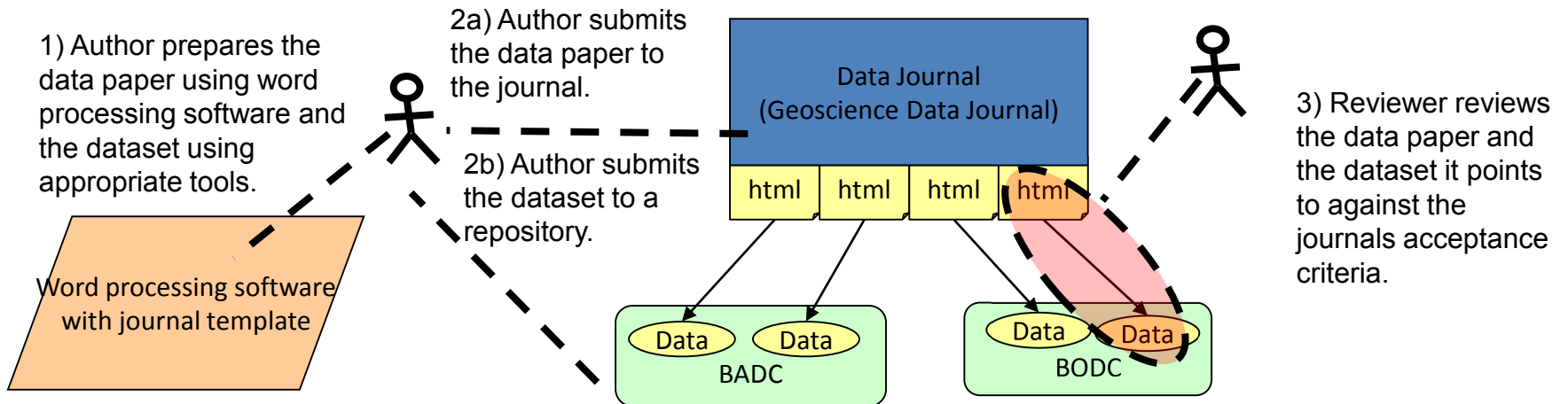


How we publish data

The traditional online journal model



Overlay journal model for publishing data



Peer Review

GDJ Reviewers consider three sets of questions

Review I – Data description document

1. Is the method used to create the data of a high scientific standard?
2. Is enough information provided (in metadata also) to enable the data to be re-used or the experiment to be repeated?
3. Does the document provide a comprehensive description of all the data that is there?
4. Does the data make an important and unique contribution to the meteorological sciences?
5. What range of applications to meteorological sciences does it have?
6. Are all contributors and existing work acknowledged?

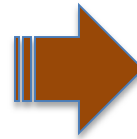


Peer Review

GDJ Reviewers consider three sets of questions

Review II – Metadata

7. Does the metadata establish the ownership of the data fairly?
8. Is enough information provided (in data description document also) to enable the data to be re-used or the experiment to be repeated?
9. Are the data present as described, and accessible from a registered repository using the software provided?



Overlaps with repository appraisal, curation processes...and trust certification?

Peer Review

GDJ Reviewers consider three sets of questions

Review III – Data themselves

10. Are the data easily readable, e.g. across different platforms such as Linux Mac and Windows?
11. Are the data of high quality e.g. are error limits and quality statements adequate to assess fitness for purpose, is spatial or temporal coverage good enough to make the data useable?
12. Are the data values physically possible and plausible?
13. Are there missing data that might compromise its usefulness?



Overlaps with repository appraisal, curation processes...and trust certification?

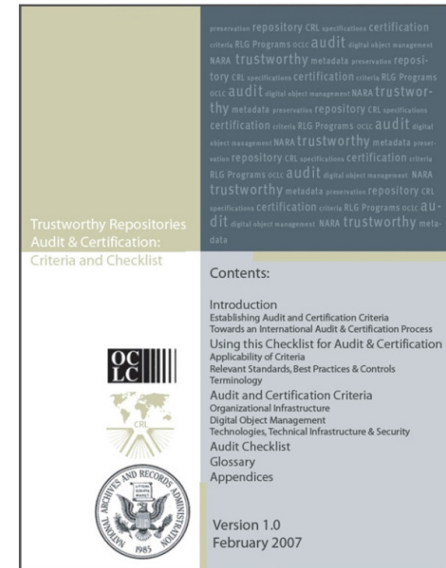
Repository accreditation schemes



European Framework for Audit and Certification of Digital Repositories.

Three levels, in increasing trustworthiness:

1. **Basic** Certification is granted to repositories which obtain DSA (Data Seal of Approval) certification;
2. **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;
3. **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."



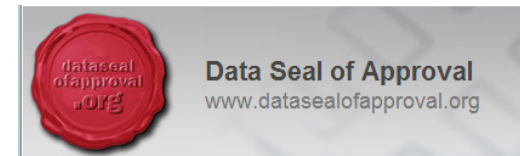
Recommendation for Space Data System Practices

AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES

RECOMMENDED PRACTICE

CCSDS 652.0-M-1

MAGENTA BOOK
September 2011



Repository accreditation – IDCC workshop

Link between data paper and dataset is crucial!

- How can data journal editors know a repository is trustworthy
- How can repositories prove they're trustworthy
- What does “trustworthy” mean for data journal peer review?

What guidelines can journals use?

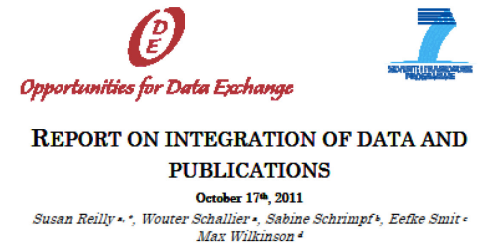
- General, cross-disciplinary and concrete
- How far do certification standards help



IDCC Workshop background

PREPARDE Workshop, Amsterdam 17 Jan. 2013

- Research Data Alliance - Working Group on Repository Accreditation
<http://forum.rd-alliance.org/viewtopic.php?f=3&t=31>
- Previous work on integrating data and publications e.g. DRIVER project and Opportunities for Data Exchange report
- Innovation in data integration
E.g. PANGAEA – Elsevier since 2010
- New data journals e.g. Journal of Open Psychology Data (Ubiquity Press, DANS)



* LIBER – Association of European Research Libraries, Koninklijke Bibliotheek, National Library Of The Netherlands, Po Box 90407, 2509 Lk The Hague, The Netherlands

† Deutsche Nationalbibliothek Informationstechnik, Adickesallee 1, D-60322 Frankfurt am Main, Germany


* The International Association of STM Publishers, Prama House, 267 Banbury Road, Oxford OX2 7HT, United Kingdom

* The British Library, 96 Euston Road, LONDON NW1 2DB, United Kingdom

* Corresponding author: Susan.Reilly@KB.nl

Abstract

Scholarly communication is the foundation of modern research where empirical evidence is interpreted and communicated as published hypothesis driven research. Many current and recent reports highlight the impact of advancing technology on modern research and consequences this has on scholarly communication. As part of the ODE project this report sought to coalesce current thought and opinions from numerous and diverse sources to reveal opportunities for supporting a more connected and integrated scholarly record. Four perspectives were considered, those of the Researcher who generates or reuses primary data, Publishers who provide the mechanisms to communicate research activities and Libraries & Data centers who maintain and preserve the evidence that underpins scholarly communication and the published record. This report finds the landscape fragmented and complex where competing interests can sometimes confuse and confound requirements, needs and expectations. Equally the report identifies clear opportunity for all stakeholders to directly enable a more joined up and vital scholarly record of modern research.

 This work is licensed under a Creative Commons Attribution 3.0 Unported License

Workshop perspectives

36 Participants – range of roles

Data Centres - UKDA, PANGAEA,
BADC

Learned Society - Royal Society
Chemistry

Publisher - Elsevier

Institutions - UK, US, De, Aus,
NL, Ch.

National Libraries & Orgs -
STM Assoc. DANS (NL), NRF
(SA), BL, DCC (UK)

Common Ground

- Data journals offer **reuse and citation**

But a passing phase?

- Data journals offer **credit** to data managers
- Certification yes, it offers journals some assurances
- Collaboration key as roles & infrastructure evolve

For data publication “trust” means...

What certification standards say it is...

Collections policy

Active curation & mgmt

Long-term preservation plans

Persistent Links

Landing pages

Continuity plan

Support for multi- stage review

Repository – QA, appraisal

Peer – open or closed

User – e.g. DANS study

Journals can plan how to integrate more data into article

Don't have to look at process detail for each dataset reviewed

Data centres can support policy compliance – track outputs against grants (e.g. IDEA Data Compliance Reporting Tool) or data sharing statements

Cloudier issues

How do repository accreditation and data quality relate to each other?

What about quality of service to depositors, users?

Researchers' and other stakeholder roles ...
e.g. advocacy, tool support to gather provenance info for publication earlier?

Repository directories – informing decisions on trust?



Indicators of repository value...not covered in certification?

- Funding
- Community acceptance
- Alt-metrics – access and reuse metrics

Service level agreements, memorandums of understanding may better meet some needs than certification

Draft guidelines for journal editors

For data publication, repositories must:

1. Ensure persistence and stability of published datasets
2. Have a clear and public indication to preserve the data or have responsibility for providing access to the data over the long term
3. Assign globally unique persistent IDs to the published datasets and maintain all URLs associated with those IDs
4. Provide persistent, actionable links to enable citations to data
5. Ensure that data will be accessible (open data, or info on license terms)
6. Actively manage and curate the data in their archive
7. Appropriate, formal succession plan, contingency plans/ escrow in case cease
8. Provide info on numbers of deposits and frequency of user access

Draft guidelines for journal editors

Repositories can 'prove' capabilities to provide persistent access by...

1. Certification on any of 3 levels in TrustedDigitalRepository.eu
2. Regular or network membership of ICSU World Data System
3. Data Centre accreditation via MEDIN
4. Contractual arrangement with DataCite managing agent to mint DOIs
5. Operate using the OAIS reference model
6. Clear intent in mission statement, institutional data mgmt policy, data preservation plan, collections policy
7. Evidence of community take-up e.g. user numbers, service level agreements, partnership agreements with well established journals, a learned society or equivalent body.

Use directory e.g. Re3data for reference on some of above.

Landing page requirements

Permanent IDs for the dataset must resolve to a publicly accessible landing page which must:

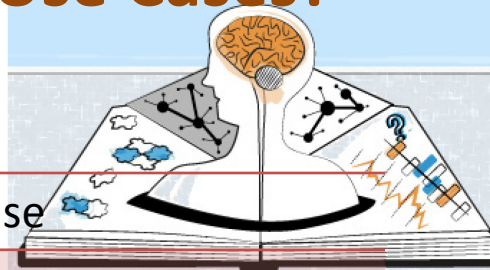
- be open and human readable (can also be provided in a format which is machine readable)
- describe the data object and include metadata and permanent identifier
- be maintained, even if the data is no longer available.

Metadata:

- Must be human readable, where possible machine readable (e.g. DataCite metadata schema)
- Freely available for discovery purposes
- Repo must develop and implement suitable quality control measures to ensure the metadata is correct



Social Science Use Cases?



The Journal of Open Psychology Data is launching soon

Submit a paper

Data centres	increase reuse
Funders, data centres, researchers, learned societies	improve transparency
Data centres, researchers, learned societies, institutions	Improve visibility
Data managers	Publication route, get credit
Researchers	Provide snapshot of rich content, sensitive data
Reusers	Support meta-analysis Mine structured description Visualisation

Thank you

And please! Tell us what you think

data-publication@jiscmail.ac.uk

sarah.callaghan@stfc.ac.uk

Project website: <http://proj.badc.rl.ac.uk/preparde/wiki>

Project blog: <http://proj.badc.rl.ac.uk/preparde/blog>



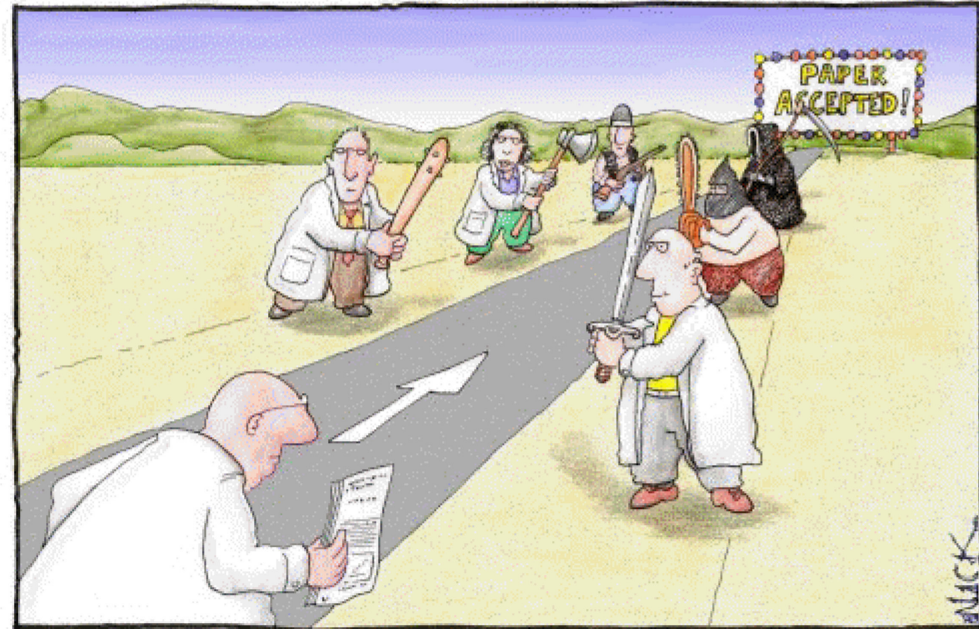
Peer-review of data

Summary Recommendations from
Workshop at the British Library, 11 March
2013

Workshop attendees included funders,
publishers, repository managers and
other interested parties.

Draft recommendations put up for
discussion and feedback from audience
captured.

Feedback from the community still
welcome!



Most scientists regarded the new streamlined
peer-review process as 'quite an improvement.'

<http://libguides.luc.edu/content.php?pid=5464&sid=164619>

Connecting data review with data management planning

1. All research funders should at least require a “data sharing plan” as part of all funding proposals, and if a submitted data sharing plan is inadequate, appropriate amendments should be proposed.
2. Research organisations should manage research data according to recognised standards, providing relevant assurance to funders so that additional technical requirements do not need to be assessed as part of the funding application peer review. (Additional note: Research organisations need to provide adequate technical capacity to support the management of the data that the researchers generate.)
3. Research organisations and funders should ensure that adequate funding is available within an award to encourage good data management practice.
4. Data sharing plans should indicate how the data can and will be shared and publishers should refuse to publish papers which do not clearly indicate how underlying data can be accessed, where appropriate.

Connecting scientific, technical review and curation

1. Articles and their underlying data or metadata (by the same or other authors) should be multi-directionally linked, with appropriate management for data versioning.
2. Journal editors should check data repository ingest policies to avoid duplication of effort , but provide further technical review of important aspects of the data where needed. (Additional note: A map of ingest/curation policies of the different repositories should be generated.)
3. If there is a practical/technical issue with data access (e.g. files don't open or exist), then the journal should inform the repository of the issue. If there is a scientific issue with the data, then the journal should inform the author in the first instance; if the author does not respond adequately to serious issues, then the journal should inform the institution who should take the appropriate action. Repositories should have a clear policy in place to deal with any feedback.

Connecting data review with article review

1. For all articles where the underlying data is being submitted, authors need to provide adequate methods and software/infrastructure information as part of their article. Publishers of these articles should have a clear data peer review process for authors and referees.
2. Publishers should provide simple and, where appropriate, discipline-specific data review (technical and scientific) checklists as basic guidance for reviewers.
3. Authors should clearly state the location of the underlying data. Publishers should provide a list of known trusted repositories or, if necessary, provide advice to authors and reviewers of alternative suitable repositories for the storage of their data.
4. For data peer review, the authors (and journal) should ensure that the data underpinning the publication, and any tools required to view it, should be fully accessible to the referee. The referees and the journal need to then ensure appropriate access is in place following publication.
5. Repositories need to provide clear terms and conditions for access, and ensure that datasets have permanent and unique identifiers.