Guidelines for Data Publication: Outputs from the PREPARDE project (public)

SPM1.38, Fri 12 April, 12:15–13:15, R3

Sarah Callaghan, Fiona Murphy, Jonathan Tedds, John Kunze, Rebecca Lawrence, Matthew S. Mayernik, Angus Whyte, Timothy Roberts and the PREPARDE project team

#preparde

sarah.callaghan@stfc.ac.uk @sorcha_ni

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Please! Tell us what you think

Always happy to get input from others!

#preparde

sarah.callaghan@stfc.ac.uk

@sorcha_ni

data-publication@jiscmail.ac.uk

Image Credit: http://bit.ly/9H4qBX

Workshop on cross-linking between data centres and publishers 30th April 2013 at Rutherford Appleton Laboratory, UK

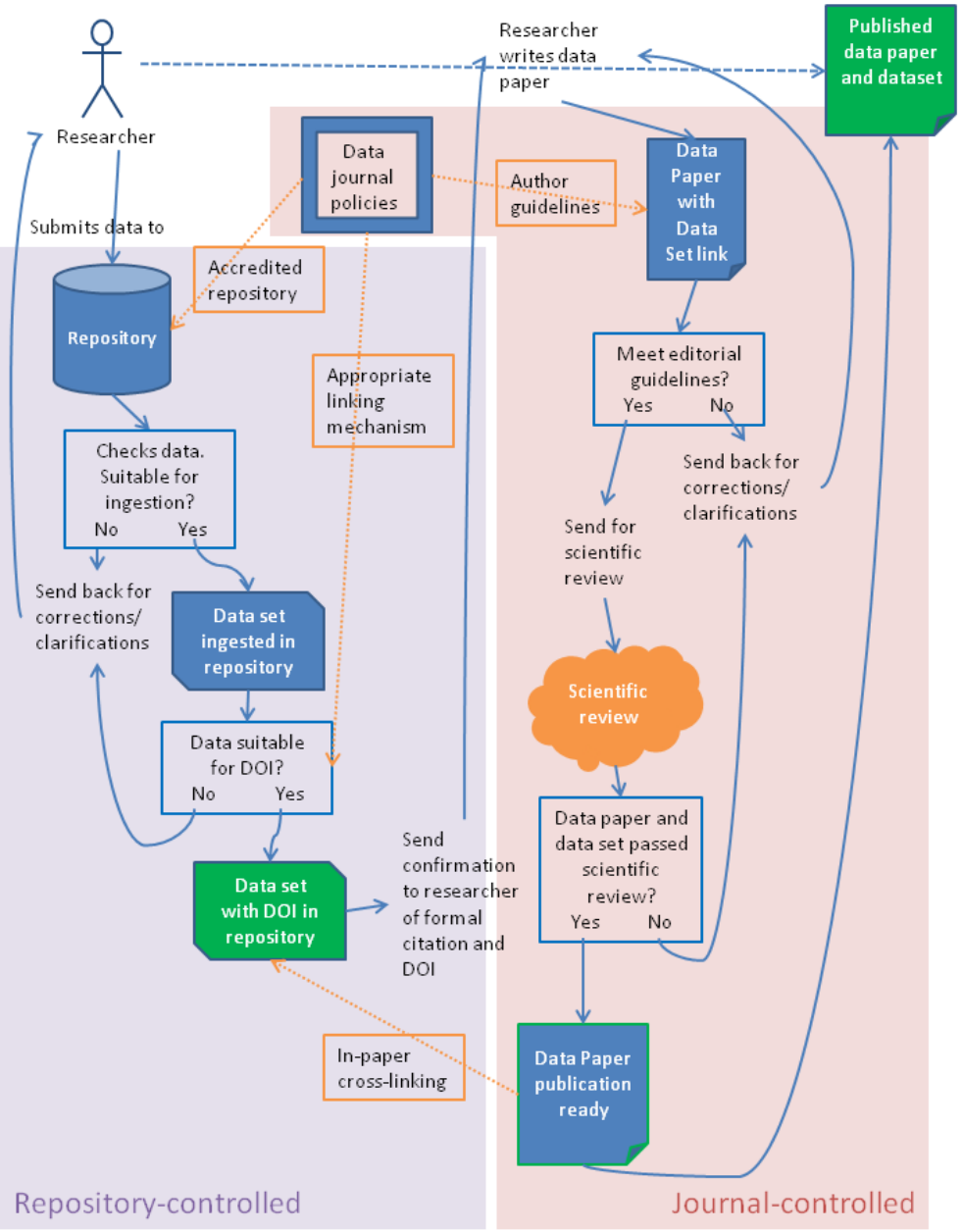
Project website: http://proj.badc.rl.ac.uk/preparde/wiki
Project blog: http://proj.badc.rl.ac.uk/preparde/blog

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

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

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH



PREPARDE topics and aims

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.

How we publish data

The traditional online journal model

1) Author prepares the paper using word processing software.

Word processing software with journal template

2) Author submits the paper as a PDF/Word file.

A Journal (Any online journal system)

PDF PDF PDF PDF PDF

Data

3) Reviewer reviews the PDF file against the journal's acceptance criteria.

Overlay journal model for publishing data

1) Author prepares the data paper using word processing software and the dataset using appropriate tools.

Word processing software with journal template

2a) Author submits the data paper to the journal.

2b) Author submits the dataset to a repository.

Data Journal (Geoscience Data Journal)

html html html html

Data Data BADC

Data Data BODC

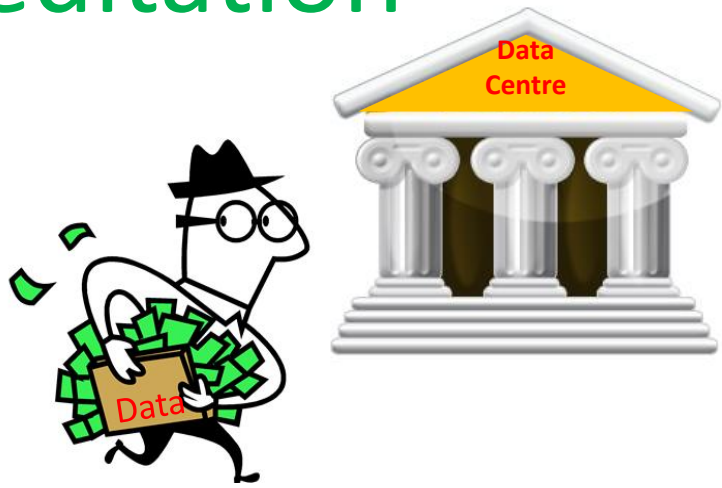
3) Reviewer reviews the data paper and the dataset it points to against the journals acceptance criteria.

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library

F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Repository accreditation

- Link between data paper and dataset is crucial!
  - How do data journal editors know a repository is trustworthy?
  - How can repositories prove they're trustworthy?



- What makes a repository trustworthy?
  - Many things: mission, processes, expertise, workflows, history, systems, documentation, …
  - Assessing trustworthiness requires assessing the entire repository workflow

And what does "trustworthy" mean, when you get right down to it?

- ***PREPARDE / IDCC13 Workshop – report in draft***
- Peer review of data is implicitly peer review of repository

Document at: http://bit.ly/ZhYHZl
Feedback to: https://www.jiscmail.ac.uk/DATA-PUBLICATION

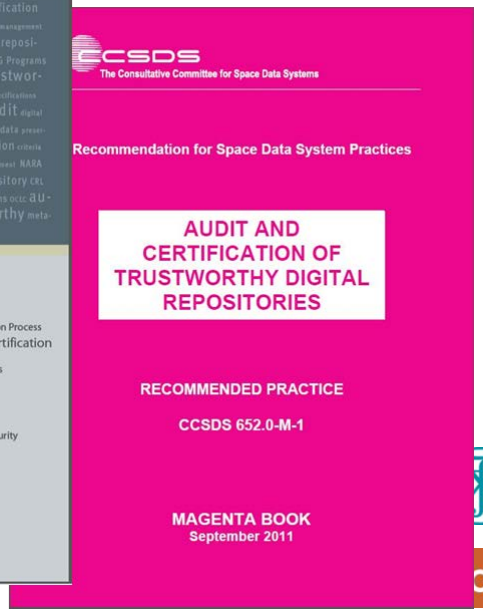
JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Repository accreditation schemes:

DataCite

Helping you to find, access, and reuse data

ICSU WORLD DATA SYSTEM



European Framework for Audit and Certification of Digital Repositories. "The framework will consist of a sequence of three levels, in increasing trustworthiness:

- Basic Certification is granted to repositories which obtain DSA (Data Seal of Approval) 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."

Membership of ICSU World Data System (criteria at http://icsu-wds.org/images/files/WDS_Certification_Summary_11_June_2012.pdf

Contractual agreement with DataCite to mint DOIs

WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library

NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

For data publication, repositories must:

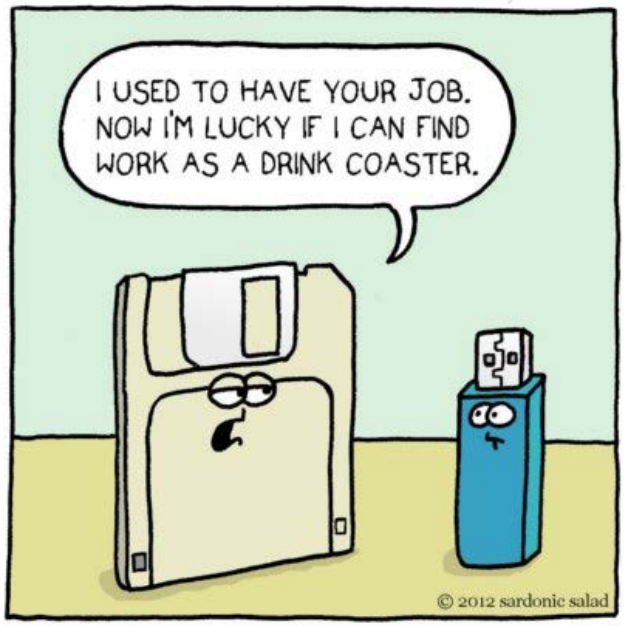
- Ensure the persistence and stability of published datasets.
- Have a clear and public indication to preserve the data or have responsibility for providing access to the data over the long term.
- Assign permanent IDs to the published datasets and maintain all URLs associated with the permanent IDs.
- Provide persistent, actionable links to enable citations to data held in their archive.
- Ensure that data will be accessible (either as open data, or provide information on licensing terms).
- Actively manage and curate the data in their archive.
- Have an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.
- Provide information on numbers of data packages or files deposited and how frequently these are accessed by repository users.

http://lolcatzencyclopedia.files.wordpress.com/2011/02/lolcat___computer_eating_by_tenkyougan1.jpg

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

How to prove data persistence:

- Regular or network membership of WDS
- Data Seal of Approval/certification under European Framework for Audit and Certification o Digital Repositories.
- Repository operates using the OAIS reference model.
- Contractual arrangement with a DataCite managing agent for the purposes of minting DOIs
- A clear intention in a mission statement or institutional data management policy, supported by a formal data preservation plan or collections policy, and evidence of community take-up such as an operational service level agreement, partnership agreement with well-established journals, a learned society or equivalent body.



http://sardonicsalad.com/?p=667

Landing page requirements

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

- 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
- must be maintained, even if the data is no longer available.

Metadata:

- Metadata about the dataset must be provided in human readable form, and when possible standardized machine readable formats (for example: DataCite metadata schema http://schema.datacite.org)
- Metadata must be made freely available for discovery purposes and must be provided on the landing page.
- Repositories should develop and implement suitable quality control measures to ensure the metadata is correct.

Peer-review of data

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

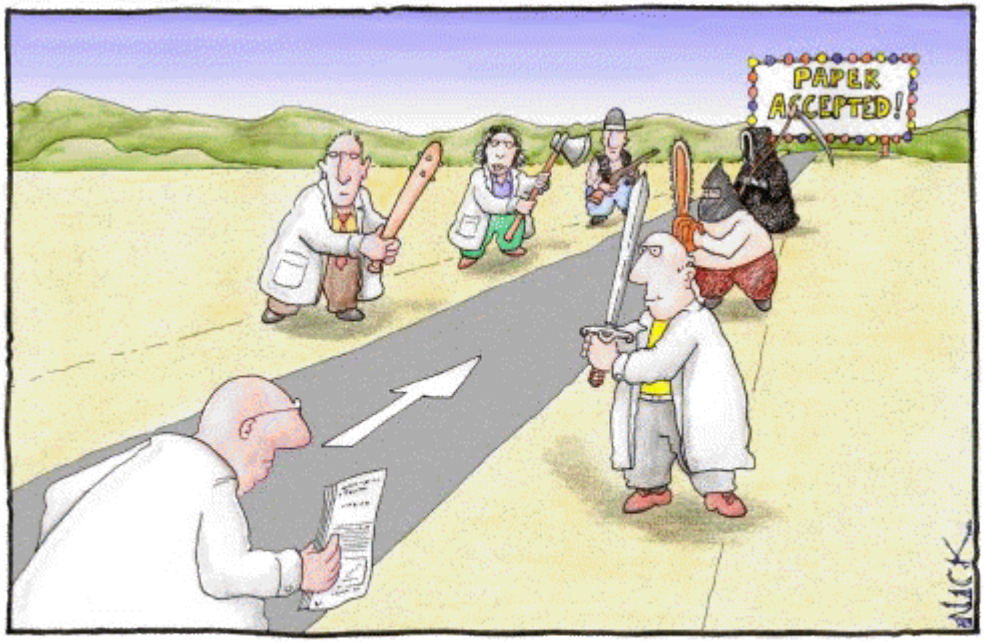
Workshop attendees included funders, publishers, repository managers and other interested parties.

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

Feedback from the community still welcome!

Document at: http://bit.ly/DataPRforComment

Feedback to: https://www.jiscmail.ac.uk/DATA-PUBLICATION



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

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

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

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.


JISC

University of
Leicester

British Atmospheric
Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

WILEY-
BLACKWELL

University of
Reading

DCC

University of California
CDL
California Digital Library

F1000
FACULTY of 1000
POST-PUBLICATION PEER REVIEW

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

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

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000 FACULTY of 1000 POST-PUBLICATION PEER REVIEW NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

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

- Partnership formed between Royal Meteorological Society and academic publishers Wiley Blackwell 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, giving details of its 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.

RMetS
Royal Meteorological Society

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL Univ Rea WILEY-BLACKWELL DL a Digital Library

F1000 FACULTY of 1000 POST-PUBLICATION PEER REVIEW NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH