DATA PUBLICATION:
INTEGRATION BETWEEN REPOSITORIES AND JOURNALS

Rebecca Lawrence, PhD
Publisher, *F1000Research*

rebecca.lawrence@f1000.com
http://f1000research.com

@f1000research

F1000Research

OVERVIEW

- Why cite and publish data?
- PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences Project
  - Journal and repository workflows
  - Repository accreditation
  - Data review
- *F1000Research*: Beyond data-only journals
- Data Publication: Key challenges

WHY CITE AND PUBLISH DATA?

- Increasing pressure from government to make data from publicly funded research available for free.
- Public want to know what the scientists are doing.
- Research funders want reassurance that they're getting value for money.
- Scientists want attribution and credit for their work.
- Extra incentive for scientists to submit their data to data centres in appropriate formats and with full metadata.
- Allows the wider research community to find and use datasets, and investigate the quality of the data.

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
- F1000 Research Ltd

Funder: JISC Managing Research Data (MRD) Programme

Project Lead: Dr Jonathan Tedds (University of Leicester)
Project Manager: Dr Sarah Callaghan (BADC)
Project dates: 1 July 2012 to 31 June 2013

PREPARDE TOPICS

- Initial focus is launch of a data journal in earth sciences: *Geoscience Data Journal.*
- *F1000Research* broadening project to the life sciences.

3 main areas of interest:

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

Responsibilities divided between:

- Repository-controlled processes and workflows.
- Journal-controlled processes and workflows.

1. JOURNAL & REPOSITORY WORKFLOWS

Data repository workflows:

- Data centre and journal workflows captured:
  - Workflows are very varied.
  - Can have multiple workflows in the same data centre, depending on interactions with external sources (“Engaged submitter”/ “Data dumper” / “Third-party requester”).

Journal workflows:

- Aim is to minimise effort needed to submit a data paper by taking advantage of already submitted metadata.
- Sharing metadata also ensures that additions/corrections made in one location get propagated through to the others.

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

2. 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?

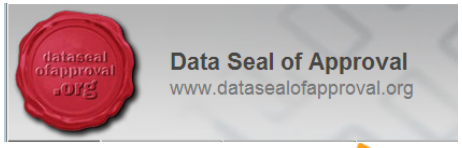
What makes a repository trustworthy?

- Mission, processes, expertise, workflows, history, systems, documentation …
- Assessing trustworthiness requires assessing the entire repository workflow.

There are many repository accreditation schemes. Look at everything relating to running a repository. Data for publication needs to:

- Be persistent.
- Be permanently identified.
- Be provided with a landing page.
- Have standard publication metadata.
- Have accessibility/licensing information.

Workshop in Jan 2013: Report in draft.



3. DATA REVIEW

- Workshop at the British Library, 11 March 2013.
- Workshop attendees included researchers, funders, institutions, repositories and publishers (also now a Research Data Alliance Working Group).
- Co-organised by:
  - Geraldine Stoneham-Clement (MRC)
  - Elizabeth Newbold (British Library)
  - Jonathan Tedds (University of Leicester; PREPARDE JISC-MRD project)
  - Rebecca Lawrence (*F1000Research*)
- Aim**:** To generate recommendations for each stakeholder group covering 3 fundamental issues:
  1. Connecting data review with data management planning.
  2. Connecting scientific, technical review and curation.
  3. Connecting data review with article review.

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

Feedback to: https://www.jiscmail.ac.uk/DATA-PUBLICATION
rebecca.lawrence@f1000.com

MOVING BEYOND DATA-ONLY JOURNALS

Datasets are rarely published alongside traditional articles

Some journals (e.g. *J Neurosci*) actively discourage publication of data

Without data publication:

- Reader must take it on faith that data were collected and analysed correctly
- Often difficult to get data from authors, limiting use and reuse
- Replication almost impossible

And even with publication:

- Data often unusable. In supplementary files, in obscure formats and poorly structured.
- Licences often limit computational mining and reuse.

F1000Research*: Data submission is *mandatory

F1000RESEARCH: MAKING DATA PUBLICATION MANDATORY

- Almost none of our authors realised they needed to provide their underlying data for publication.
- A small number raised the usual concerns:
  - Wanting to publish other papers from the datasets.
  - Don't want others to scoop the work until finished own data analysis.
  - Too much confidential data.
  - Too time-consuming to explain data to potential readers/users.

Despite this, **all authors have provided their datasets**

Why? The main reason:
Publishing your data provides you with **priority on the data**

F1000RESEARCH: DATA PRE-PUBLICATION CHECKS

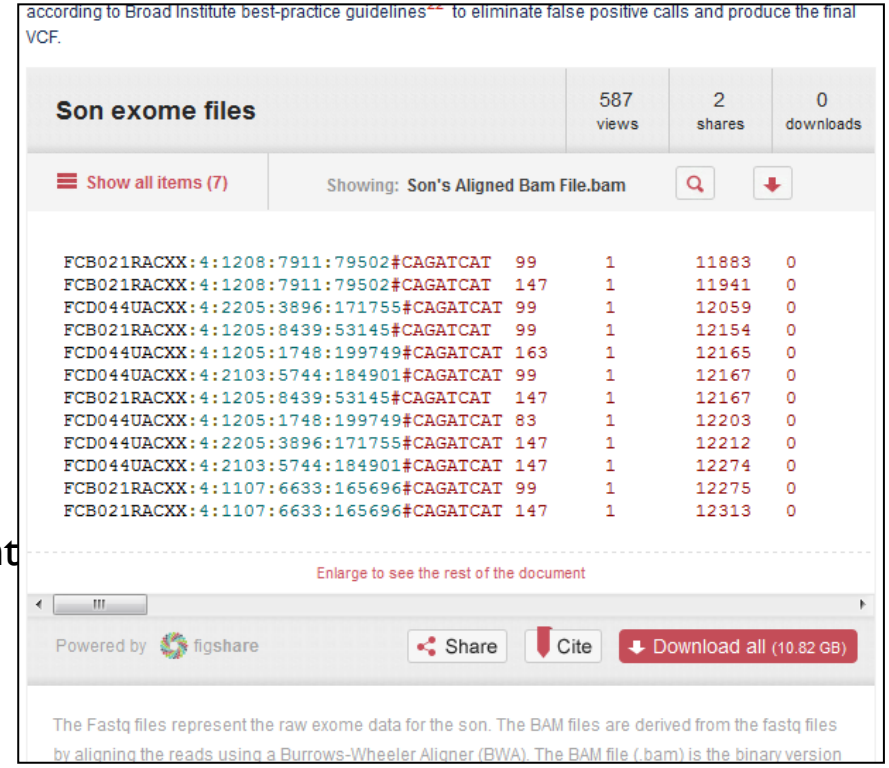
At *F1000Research* we undertake a pre-publication data review:

- Are there any subject-specific repositories the data should be placed into?
- Are the formats appropriate?
- Is the layout understandable? Is labelling clear?
- Do we have adequate data?
- Do we have adequate protocol information about how the data was generated?
- If no existing repository or suitable alternative, we work with **figshare.**

F1000RESEARCH: EMBED WIDGETS

Collaboration with **figshare**:

- Users can view the data without leaving the article
- Figshare provides viewers for data files
- Users can preview large datasets before deciding whether to download
- Usage information provided.
- Datasets get legends and DOIs: Independent citation.



F1000RESEARCH: DATA PEER REVIEW

Referees are asked to check:

- Is the method used appropriate for the scientific question being asked?
- Has enough information been provided to be able to replicate the experiment?
- Are the data in a useable format/structure?
- Are stated data limitations and possible sources of error appropriately described?
- Does the data 'look' OK (optional; e.g. microarray data)?

The ultimate referee: Reuse!

DATA PUBLICATION: KEY CHALLENGES

- Encouraging the accreditation of repositories.
- Developing stronger links between repositories and journals, in both directions: workflows and review outputs.
- Stronger ‘carrots’ for data sharing, such as mandatory data release on publication.
- Development of better credit systems for the sharing, curation and publication of data.

Thank you!

rebecca.lawrence@f1000.com

@f1000research