WDS/RDA Publishing Data Interest Group
Workflows Working Group
Draft Case Statement

WG Charter

Objectives
What are possible workflows for publishing data? What are the experiences gained so far?
The objectives of this Working Group are to investigate the range of possible workflows for publishing data, in particular reporting on:

- Investigate current workflows for archiving and publishing data
- The role of QA/QC and peer-review in the publication process
- The role of science publishers and journals in the data publication process
- Identify key barriers for identified use cases

In particular we will build on the work carried out in the UK Jisc funded PREPARDE[1] project which provided reports on journal workflows from author submission of datasets and papers, through review to publication. The project also investigated data repository workflows from ingestion of data, through data centre technical review, to DOI assignment to dataset. While the focus was on mature examples in the Earth Sciences the reports also considered alternative paradigms which will be extended through the work of this group.

Deliverable: Provide generic workflow models for data publication across a characteristic range of use cases in different disciplines. We will provide the international research community with clear examples that may be adapted for use, in each case identifying the varying stakeholders and their different roles and responsibilities as well as the likely associated resource and cost implications.

Value Proposition

Often, there are insufficient incentives for data submission, resulting in low submission rates and even when submitted, a bare minimum of metadata. Persistence, quality control and access all enhance the possibilities for greater discoverability and re-use of research data. The development of workflows and standards for data collection, curation, storage and citation will greatly enable future research to be aware of, and build upon, work already undertaken. Reproducibility, normalised results ranges and policy recommendations will also be supported by data publication. comment

The assignment of persistent identifiers, specifically Digital Object Identifiers (DOIs), enables accurate data citation. Data publication that enables data citation can certainly be an incentive to make data accessible. Scientists are now becoming aware that Digital Object Identifiers (DOIs) offer the

[1] http://www.le.ac.uk/projects/preparde

means to easily cite their datasets and gain citation metrics.

Who will benefit:
Given that the mechanism of publishing data will greatly enhance discoverability, interoperability, re-use, transparency and accountability, the potential benefit to all the stakeholders - authors, funders, policy makers, librarians, publishers, and so forth - seems self evident.

Authors will be able to derive credit from adhering to best practice in managing and sharing data collected, funders will be able to track the research data they have funded, measure impact and guard against repetition. Researchers will be able to work faster, achieve deeper insights outside their immediate subject domain. Librarians and data centre managers become an integral part of the ecosystem through their expertise in cataloguing and metadata production. Policy makers and the public will be able to navigate the knowledge landscape with increased confidence in its veracity.

Impact:
In order to reach this point, however, some major adjustments will need to be made by all of these participants. [2] Publishing models are in transition in any case from 'pay to read' to 'pay to publish', but so far this applies to primary research output - and not to research data in its own right. New business models and mechanisms for compiling metrics, such as the newly formed Thomson Reuters Data Citation Index, are very much in their infancy and have not yet permeated the general researcher consciousness. Many researchers are inherently suspicious of calls to share data more widely. Funders, dealing with the inundation of the "Open Access" issue, and with many national agencies suffering budget freezes or cuts, are also in a state of transition.

Engagement with existing work in the area

An overview of relevant initiatives, projects, and platforms will be developed and maintained at the level of the Publishing Data Interest Group, and may be found here: https://docs.google.com/spreadsheet/ccc?key=0AoqnUFYMgSc-dGhna3ZnNHhmRkNxRzZTMFBrZUpNQ3c#gid=0

- IODE/SCOR/MBLWHOI

  The Marine Biological Laboratory/Woods Hole Oceanographic Institution (MBLWHOI) Library, the Scientific Committee on Oceanic Research (SCOR) and the International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission have assembled a team of librarians, data managers and scientists who are collaborating to identify best practices for tracking data provenance and clearly attributing credit to data collectors/providers.

  The group has developed and executed pilot projects related to two use cases: (1) data held by data centers are packaged and served in formats that can be cited and (2) data related to traditional journal articles are assigned persistent identifiers and stored in institutional repositories. IODE has a history of fostering the establishment of standards and this collaboration is building a "community" of librarians, data managers and scientists to address the data publication paradigm.

  The assignment of persistent identifiers enables accurate data citation. The

[2] See Thomson Reuters Industry Forum report 'Unlocking the Value of Research Data', July 2013 on the challenges currently inherent in the system.

MBLWHOI Library is assigning Digital Object Identifiers (DOIs) to appropriate datasets deposited in the Institutional Repository (IR), Woods Hole Open Access Server (WHOAS). We are particularly interested in working with authors to deposit datasets associated with published articles. The DOI would ideally be assigned before submission and be included in the published paper so readers can link directly to the dataset, but DOIs are also being assigned to datasets that support papers that have already been published. WHOAS metadata records link the article to the datasets and the datasets to the article. The repository is indexed in Data Citation Index.

The Published Data Library (PDL) is a project of the British Oceanographic Data Centre that provides snapshots of specially chosen datasets that are archived using rigorous version management. The publication process exposes a fixed copy of an object and then manages that copy in such a way that it may be located and referred to over an indefinite period of time. Using metadata standards adopted across NERC's Environmental Data Centres, the repository assigns DOIs obtained from the British Library to appropriate datasets.

One successful outcome of this collaborative effort includes tools and procedures developed by the MBLWHOI Library and the Biological and Chemical Oceanography Data Management Office (BCO-DMO) that automate the ingestion of metadata from BCO-DMO for deposit with a copy of each data set into the IR, WHOAS. The system also incorporates functionality for BCO-DMO to request a DOI from the Library. This partnership allows the Library to work with a trusted data repository to ensure high quality data while the data repository utilizes library services and is assured that a permanent archived copy of the data is associated with the persistent DOI.

The project team has also produced the “Ocean Data Publication Cookbook” available at www.iode.org/mg64

- The Inter-university Consortium for Political and Social Research (ICPSR), a repository of social and behavioral science research data established in 1962, has a documented workflow that tracks data from deposit through curation to publication when the data become discoverable and accessible on the ICPSR Web site. ICPSR has over 8000 studies comprising about 65,000 datasets and has been providing data citations since 1990 and DOIs since 2008. The archive includes a smaller collection of replication datasets intended to reproduce findings in publications. ICPSR also provides two-way linking between data in the holdings and citations to publications. The ICPSR Bibliography of Data-Related Literature includes over 60,000 citations to publications that reference ICPSR data. Like the WHOAS data described above, the ICPSR repository is indexed in the Thomson Reuters Data Citation Index.
- PANGAEA (Michael)
- Springer (Johanna)
- Geoscience Data Journal/Wiley (Fiona/Jonathan)
- Astronomy (Jonathan)
- Genome community (EBI, Jo McEntyre)
- Other in the humanities? The Language Archive?

Work Plan

General approach:

- Use Cases
- Test bed

Steps to take

Adoption Plan

1. **The form and description of final deliverables of the WG**
   Deliverable: Provide generic workflow models for data publication
   a. Roles & steps of data publication, scope of data (science fields)
   b. Investigate current workflows for ingest, archiving, and publishing data
   c. The role of QA/QC and peer-review in the publication process
   d. The role of science publishers and journals in the data publication process
   e. Recommendations (essentials for workflows)
2. **The form and description of milestones and intermediate documents, code or other deliverables that will be developed during the course of the WG's work**
   a. Investigate current practice
   b. Gap analysis and general requirements
   c. Recommendations to fill the gap
   d. Define clear roles for Data Centres, Publishers, Funders
3. **A description of the WG's mode and frequency of operation**
   TBC
4. **A description of how the WG plans to develop consensus, address conflicts, stay on track and within scope, and move forward during operation**
   TBC
5. **A description of the WG's planned approach to broader community engagement and participation**
   TBC

Membership

- Jonathan Tedds, (UK, University of Leicester) **[Chair]**
- Kim Finney, (Australia, AADC)
- John Helly, (US, UCSD)
- Hylke Koers, (The Netherlands, Elsevier)
- Fiona Murphy, (UK, Wiley-Blackwell)
- Amy Nurnberger, (US, Columbia University Libraries)
- Lisa Raymond, (US, Library Woods Hole Oceanographic Institution)
- Johanna Schwarz, (Germany, Springer)
- Mary Vardigan, (USA, ICPSR)
- Eva Zanzerkia, (US, NSF)