

# Publishing Data Workflows

---

*Chairs:*

Theodora Bloom (BMJ)

Sünje Dallmeier-Tiessen (CERN)

Elizabeth Newbold (British Library)



# Agenda

- Our objectives
- Where are we now?
- Next steps!

# Background and Motivation

- Policy pressure vs. lack of incentives for researchers
- Only a small fraction of research data is preserved and shared
  - often with a bare minimum of metadata
- We believe this is often due to the lack of “established” or “trusted” services and workflows

**But there are established or emerging workflows!**

- Usually in selected disciplines e.g. Earth Sciences
- Some provide credit via citation mechanisms

# Relevance

- Policy pressure vs. lack of incentives for researchers
- Having information about workflows is crucial
  - for researchers
  - and the people/stakeholders supporting themto understand the options available to practice open science
- That's why we propose to study and test workflows that allow
  - efficient and reliable reuse of research data
  - to enhance the possibilities for greater discoverabilityacross disciplines and stakeholder groups

# The working group members (currently)

- Theodora Bloom (BMJ) [CO-CHAIR]
- Sünje Dallmeier-Tiessen (Switzerland, CERN) [CO-CHAIR]
- Elizabeth Newbold (BL) [CO-CHAIR]
- Merce Crosas (US, Harvard University)
- Michael Diepenbroek (PANGAEA)
- Kim Finney (Australia, AADC)
- John Helly (US, UCSD)
- Hylke Koers (The Netherlands, Elsevier)
- Rebecca Lawrence (UK, F1000 Research Ltd.)
- Fiona Murphy (UK, Wiley-Blackwell)
- Amy Nurnberger (Columbia University Libraries)
- Lisa Raymond (US, Library Woods Hole Oceanographic Institution)
- Johanna Schwarz (Germany, Springer)
- Jonathan Tedds (UK, University of Leicester)
- Mary Vardigan (US, ICPSR)
- Ruth Wilson (UK, Nature)
- Eva Zanzerkia (US, NSF)
- Angus Whyte (UK, DCC)
- Brian Hole (Ubiquity Press, UK)
- Varsha Khodiyar (UK, F1000 Research Ltd.)
- And growing...

# Objectives

- To provide an analysis of a representative range of existing and emerging workflows and standards for data publishing
  - including deposit and citation
  - provide reference models, a “classification”
- To test implementations of key components for application in new workflows.
- To illustrate the benefits of the reference model to researchers and organisations

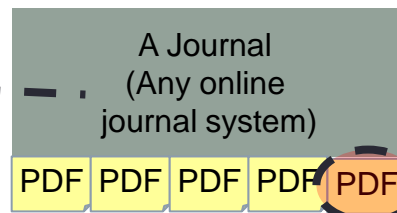
# Example: How to publish data in GDJ

## 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.



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



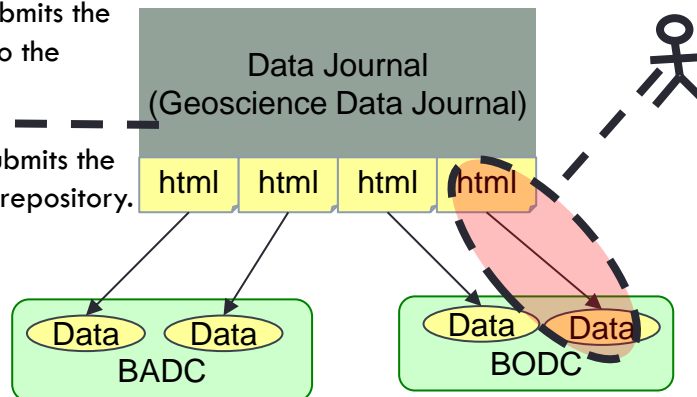
## Overlay journal model for publishing data

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.



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

# Example:



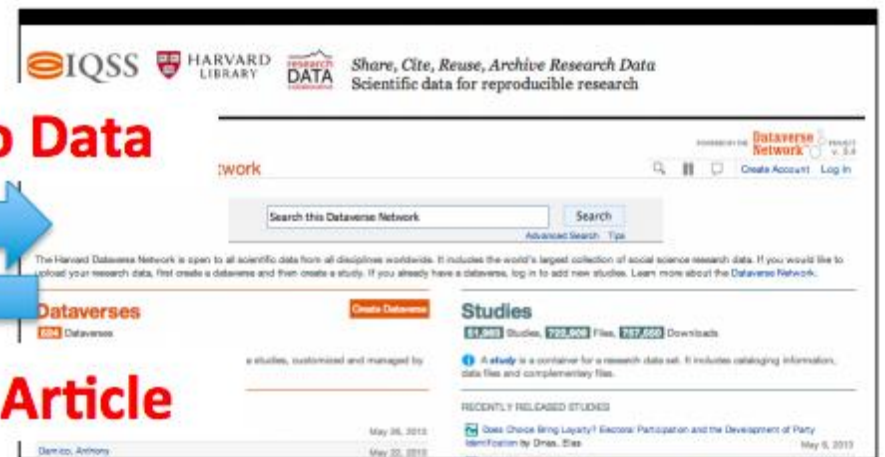
PKP  
PUBLIC  
KNOWLEDGE  
PROJECT



- Integrate PKP's Open Journal System with Dataverse
- OJS plugin for: Data + metadata + supporting files, sent via SWORD API to the Dataverse
- <http://projects.iq.harvard.edu/ojs-dvn>



OJS Journal



Harvard Dataverse Network





# More detailed work programme

- Identification of a smaller set of reference models covering a range of such workflows to include:
  - For example **when and where** QA/QC and data peer-review fit into the publishing process
  - **Who** does what and when...
  - Automated vs. “manual” processes
- Selection of key use cases and organizations in which components of a reference model can be **implemented** and tested for suitability
  - For example: dedicated data peer review
  - For example: metadata checks

# Where are we now?

- Ongoing work right now:
  - Survey planning
  - Collecting and categorizing workflow examples
- ➔ **Share your own data publishing workflows**
  - How do you choose, collaborate with or link to data repositories?
  - Do you have data peer review?
  - Do you have recommendations for data citation?

<http://tinyurl.com/mvtbrek>

RDA-WDS-F11 workflow analysis ☆

File Edit View Insert Format Data Tools Add-ons Help Last edit was made 6 days ago by anonymous

Rich text editor toolbar with icons for undo, redo, bold, italic, underline, text color, background color, bulleted list, numbered list, link, unlink, and other standard editing functions.

Some notes beforehand:

| A   | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| Some notes beforehand:  |   |   |   |   |   |   |   |   |   |
| - please feel free to add workflows you think should be part of the analysis by expanding the table to the right  |   |   |   |   |   |   |   |   |   |
| - if you think, there are categories missing (or need to be changed), please add a row and drop an email to the list to highlight the change that affects the other ongoing analysis - but please c |   |   |   |   |   |   |   |   |   |
| - please provide as many links to documentations, charts as possible to enrich the analysis   |   |   |   |   |   |   |   |   |   |

|  | Workflow name   | STFC Data centre                                    | NSIDC Data centre   | ENVRI reference model Infrastructure/Service provider (data centre/repository?) | OJS/ Dataverse Data repository  | INSPIRE Digital library                            | NPG (PubChem & Scientific Data) Publisher         | UK Data Archive/Service provider/ data centre | PREPARDE (NCAR CISL) Data centres (+ data journals)  |
|--|---|---|---|---|---|--|---|---|--|
|  | Stakeholder in charge [inst, publisher, disciplinary repository, ...] |   |   |   |   | disciplinary, international data repository        |   |   |  |
|  | Available online  | <a href="http://www.prelida">http://www.prelida</a> | <a href="ftp://sidads.colorado">ftp://sidads.colorado</a> | <a href="http://confluence.e">http://confluence.e</a>                           | <a href="http://projects.iq.harvard.edu/ojs-dvn">http://projects.iq.harvard.edu/ojs-dvn</a> & <a href="https://docs.google.com/document/d/1T-i2a4synXlthe3DCiYyALi8VYg_h2hLdJJMmd6KVVXhc/edit#">https://docs.google.com/document/d/1T-i2a4synXlthe3DCiYyALi8VYg_h2hLdJJMmd6KVVXhc/edit#</a> & <a href="https://docs.google.com/file/d/0BzeLxEN77UzoYnBLWXpodmpTLTQ">https://docs.google.com/file/d/0BzeLxEN77UzoYnBLWXpodmpTLTQ</a> | <a href="http://inspirehep.net">inspirehep.net</a> | <a href="http://www.dcc.ac">http://www.dcc.ac</a> | ODIN Deliverable [link?]                      | <a href="http://www.le.ac.uk/projects/preparde">http://www.le.ac.uk/projects/preparde</a><br><a href="http://proj.badc.rl.ac.uk/preparde/attachment/wiki/DeliverablesList/D2_1_D2_2_PREPARDE_Workflows_combined_draft1.pdf">http://proj.badc.rl.ac.uk/preparde/attachment/wiki/DeliverablesList/D2_1_D2_2_PREPARDE_Workflows_combined_draft1.pdf</a> |

# Categories we look at

- Discipline
- Function of workflow
- Includes PID assignment to data set
- PID type e.g DOI, ARK etc
- Peer review of data, (e.g. by researcher & editorial review)
- Curatorial review of metadata (e.g. by institutional or subject repository?)
- Technical review & checks (e.g. for data integrity at repository/data centre on ingest)
- Discoverability: Indexing of the data -if yes, where?
- Formats covered
- Persons/Roles involved, e.g. editor, publisher, data repository manager etc. [please list everyone you know]
- Is there a link to paper or is this standalone data
- Links to grants, usage of author PIDs
- Data Citation Facilitated
- Refers to a data life cycle

# Workflows in the current list

STFC Data centre

NSIDC Data centre

ENVRI reference model

OJS/ Dataverse Data repository

INSPIRE Digital library (HEP)

NPG (PubChem & Scientific Data) Publisher

UK Data Archive/Service

PREPARDE

Ocean Data Publication Cookbook UNESCO IOC

PURR Institutional repository

ICPSR

Edinburgh Datashare

F1000

ResearchUbiquity Press: Open Health Data Journal +...

PANGAEA - Data Publisher for Earth & Environmental Sciences

WDC Climate - Data Publisher for Climate Sciences CMIP / IPCC DDC - international project series in Climate Sciences

# How to get involved now

- Contribute to the workflow analysis
- Contribute your own workflow “walk-throughs”
- Tell us what is needed for a “successful” workflow in your institute/discipline etc.

## **... moving to implementations**

- Tell us if you are interested to learn from a specific example or are maybe considering implementing data publishing workflows
- Tell us if you have code/documentation to share