Connecting data repositories and publishers for data publication

Sarah Callaghan, Fiona Murphy, Jonathan Tedds, Rob Allan, John Kunze, Rebecca Lawrence, Matthew S. Mayernik, Angus Whyte, and the PREPARDE project team
#preparde
sarah.callaghan@stfc.ac.uk @sorcha_ni
Why link data and publications?

• Data is the foundation of science – without it we can’t test our assertions or reproduce our results

• The Internet allows us to link things to other things quickly and easily

• But there are still serious problems to address when it comes to linking data to the scientific record:
  – Data persistence
  – Data and metadata quality
  – Attribution and credit for data producers
  – ... and many more

Engraving of printer using the early Gutenberg letter press during the 15th century.
Date unknown - estimate 16th - 19th century
http://commons.wikimedia.org/wiki/File:Gutenberg_press.jpg
Historically, journals have always published data

Suber cells and mimosa leaves. Robert Hooke, *Micrographia*, 1665

The Scientific Papers of William Parsons, Third Earl of Rosse 1800-1867
But now... the Data Deluge

“the amount of data generated worldwide...is growing by 58% per year; in 2010 the world generated 1250 billion gigabytes of data”

A lot of people are creating a lot of data, and we’re only going to get more of it.

If this is a data deluge – time to start building boats!

The Digital Universe Decade – Are You Ready? IDCC White Paper, May 2010

Figure 1: The Digital Universe 2009 – 2020 Growing by a Factor of 44

2020
35 ZB*

2009
0.8 ZB*

*Zettabyte = 1 trillion gigabytes

Source: IDC Digital Universe Study, sponsored by EMC, May 2010
Using citations to link research outputs

• We already have a working method for linking between publications which is
  • commonly used
  • understood by the research community
  • used to create metrics to show how much of an impact something has (citation counts)
  • applied to digital objects (digital versions of journal articles)

• We can extend citation to other things like
  • data
  • code
  • multimedia

And the best bit is, we don’t need to teach researchers a new method of linking – they cite like they normally would!

http://www.flickr.com/photos/anton41/6588935181/
Reasons for citing and publishing data

• **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

**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
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 groundbreaking conclusions.
  • the when, how and why data was collected and what the data-product is.
**The traditional online journal model**

1) Author prepares the paper using word processing software.

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.

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 journal’s acceptance criteria.
Data paper mock-up

Dataset citation is first thing in the paper and is also included in reference list (to take advantage of citation count systems)
Example steps/workflow required for a researcher to publish a data paper

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

• Division of area of responsibilities between
  • repository controlled processes
  • journal controlled processes
Data repository workflows

- Data centre and journal workflows captured
  - Workflows are very varied! No one-size fits all method
  - Can have multiple workflows in the same data centre, depending on interactions with external sources (“Engaged submitter”/ “Data dumper” / “Third party requester”)
Repository Workflow – NCAR Comp. & Info. Systems Lab Research Data Archive (RDA)

Data Ingest

Data Preparation:
• Automated file collection.
• Check integrity of file receipts.
• Compare bytes and checksums (if available) with original data providers.

Data Preparation:
• Not ok
• Ok

Processing:
• Validate files – using software, read the full content of every file.
• Pull out metadata.
• Identify errors and metadata holes.
• Do time-series checks.
• Check metadata against internal standard/expectation.
• If necessary, filter data or fix metadata.

Check with data provider for changes to files

Embargo

Archive (Tape-based)

Metadata Database
• Spatial info
• Temporal info
• Global Change Master Directory (GCMD) keywords
• Parameters
• Format table relationships

Errors found

Contact data provider

Remote backup

Access Development Phase

Online Data (Most Demanded)

Publish Metadata – User GUIs

Distribute metadata

GCMD
NCAR CDP
BADC
... OAI-PMH

Notification to provider/user community

Notification to provider/user community
Journal workflow

- Work on comparisons and identification of cross-linking points is continuing.
- Aim is to minimise effort needed to submit data paper by taking advantage of already submitted metadata.
This is what we have to focus on for PREPARDE – demonstrate cross linking between GDJ and BADC (and maybe NCAR).

Unfortunately this direct cross-linking isn’t scaleable!

Need for off-the shelf solutions that can work across multiple research domains.
Cross-linking – the ideal situation

Registry could provide other functions as well as being an intermediary between journals and data repositories like:

- Certify data centres are “trustworthy”
- Administer linking mechanism
- Provide search and metrics functions

Disadvantages:

- Single point of failure
- Difficulty of standardisation across different research domains

Could OpenAIRE be this registry?
Do we have a start?

DataCite have standardised a set of bibliometric metadata that have to be submitted before a DOI for a dataset can be minted by a repository.

This metadata is then made openly available via the DataCite metadata search: http://search.datacite.org/ui

Given a DOI, a journal can then easily find the DOI standard metadata.

DataCite also have a content resolver http://data.datacite.org/static/index.html

What’s missing is the return link, where the journal can let the repository know that a dataset has been cited (directly or via DataCite)
## DataCite Metadata Schema

### DataCite Mandatory Properties

<table>
<thead>
<tr>
<th>ID</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Property</td>
</tr>
<tr>
<td>1</td>
<td>Identifier (with type attribute)</td>
</tr>
<tr>
<td>2</td>
<td>Creator (with name identifier attributes)</td>
</tr>
<tr>
<td>3</td>
<td>Title (with optional type attribute)</td>
</tr>
<tr>
<td>4</td>
<td>Publisher</td>
</tr>
<tr>
<td>5</td>
<td>PublicationYear</td>
</tr>
</tbody>
</table>

### DataCite Optional Properties

<table>
<thead>
<tr>
<th>ID</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Property</td>
</tr>
<tr>
<td>6</td>
<td>Subject (with schema attribute)</td>
</tr>
<tr>
<td>7</td>
<td>Contributor (with type and name identifier attributes)</td>
</tr>
<tr>
<td>8</td>
<td>Date (with type attribute)</td>
</tr>
<tr>
<td>9</td>
<td>Language</td>
</tr>
<tr>
<td>10</td>
<td>ResourceType (with description attribute)</td>
</tr>
<tr>
<td>11</td>
<td>AlternateIdentifier (with type attribute)</td>
</tr>
<tr>
<td>12</td>
<td>RelatedIdentifier (with type and relation type attributes)</td>
</tr>
<tr>
<td>13</td>
<td>Size</td>
</tr>
<tr>
<td>14</td>
<td>Format</td>
</tr>
<tr>
<td>15</td>
<td>Version</td>
</tr>
<tr>
<td>16</td>
<td>Rights</td>
</tr>
<tr>
<td>17</td>
<td>Description (with type attribute)</td>
</tr>
</tbody>
</table>
MOLES: Metadata Objects for Linking Environmental Sciences v3.4

http://proj.badc.rl.ac.uk/moles/browser/branches/V3.4/MODEL/Diagrams/MOLES3.4Summary.png
What PREPARDE is going to do

We already have a link from the GDJ data article to the data repository – thanks to the DOI.

GDJ can also pull the standard DOI metadata attached to that DOI from the DataCite metadata store.

We need to figure out a way so GDJ can inform the repository that their dataset has been cited/published – bearing in mind scaling issues!

Might have to start with a manual work-around.
Tell us what you think

Workshop on cross-linking between data centres and publishers planned for May 2013 at Rutherford Appleton Laboratory, UK

Workshop on peer-review of data planned for March 2013 at the British Library

Always happy to get input from others!

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

Thanks! Any questions?

#preparde
sarah.callaghan@stfc.ac.uk
@sorcha_ni

The project is led by the University of Leicester and the support of JISC and NERC in funding the PREPARDE project is gratefully acknowledged.
Yor data organishun
is harrible

Pweez to stand by...we seemz to be ekzpeerinsing teknikul difikulteez.