PREPARDE final meeting slides

Sarah Callaghan

#preparde

sarah.callaghan@stfc.ac.uk @sorcha_ni
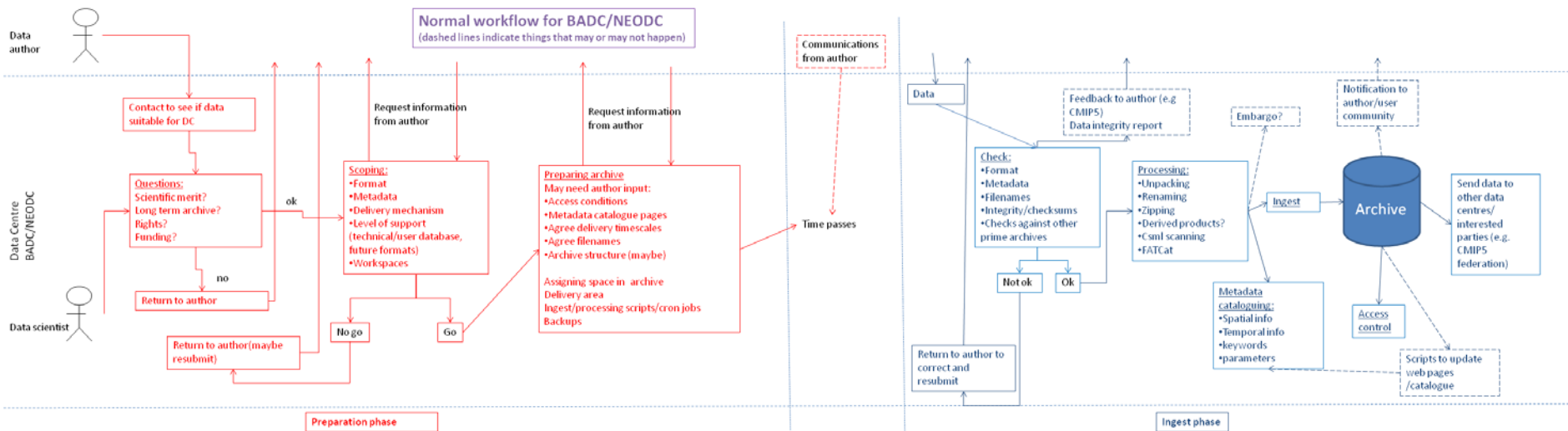
Wiley, Oxford, 23rd August 2013
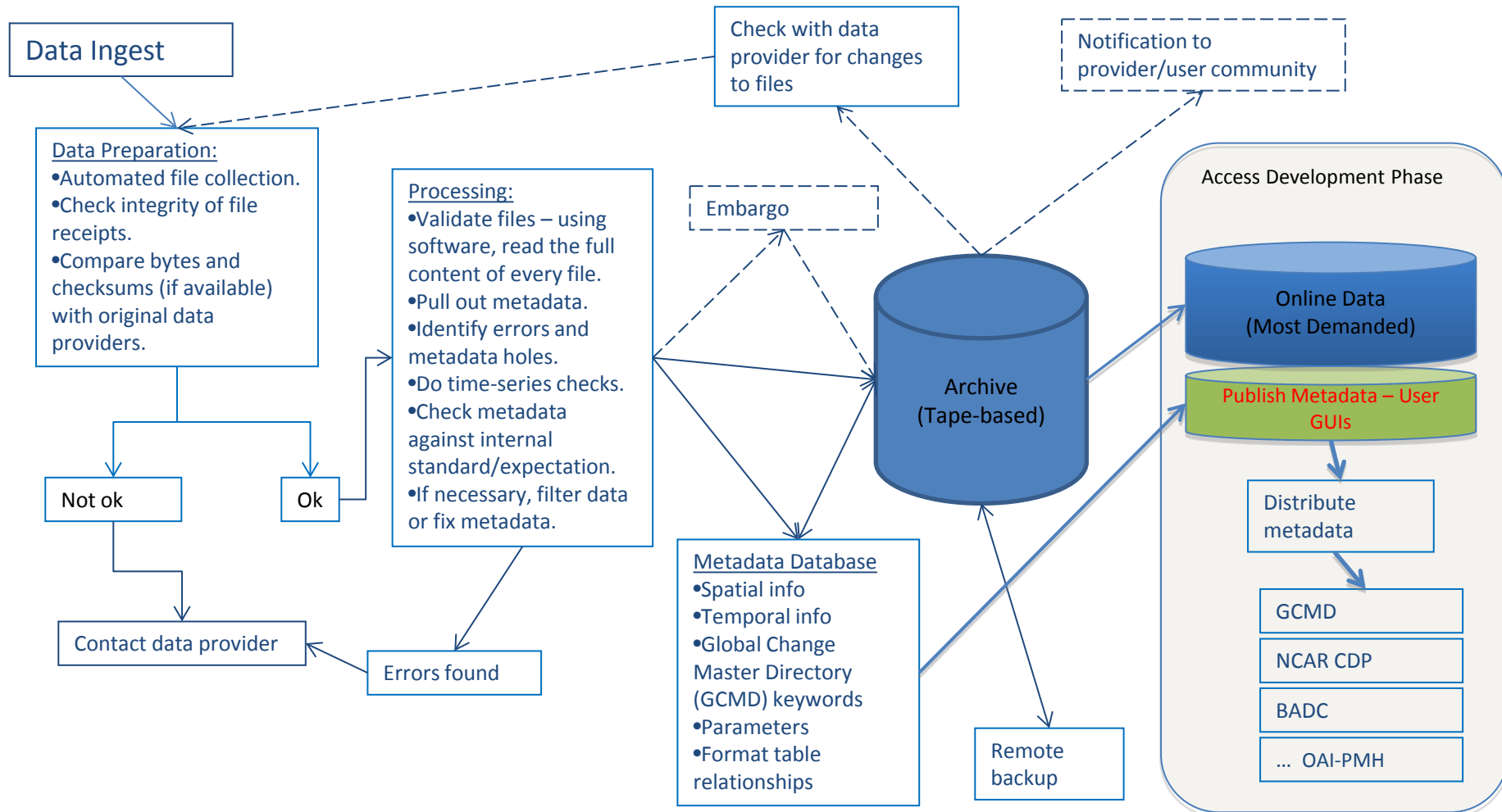
WP2 Workflows

- Data Centres
  - CEDA (broken down into type of data submitter)
  - NCAR Earth Observing Laboratory (EOL): Computing, Data, and Software Facility
  - NCAR CISL Research Data Archive (RDA), http://rda.ucar.edu/
  - NERC DOI minting workflow
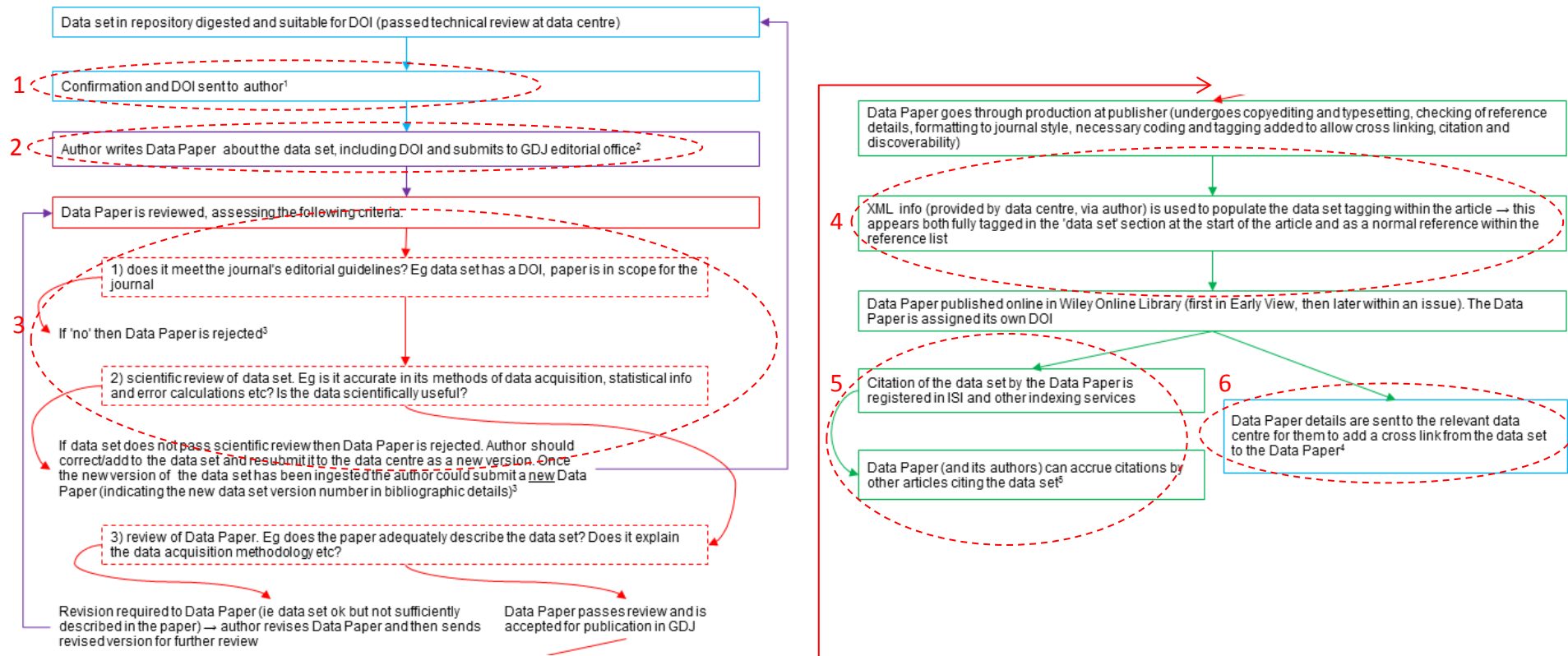
Data repository workflows

- 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")
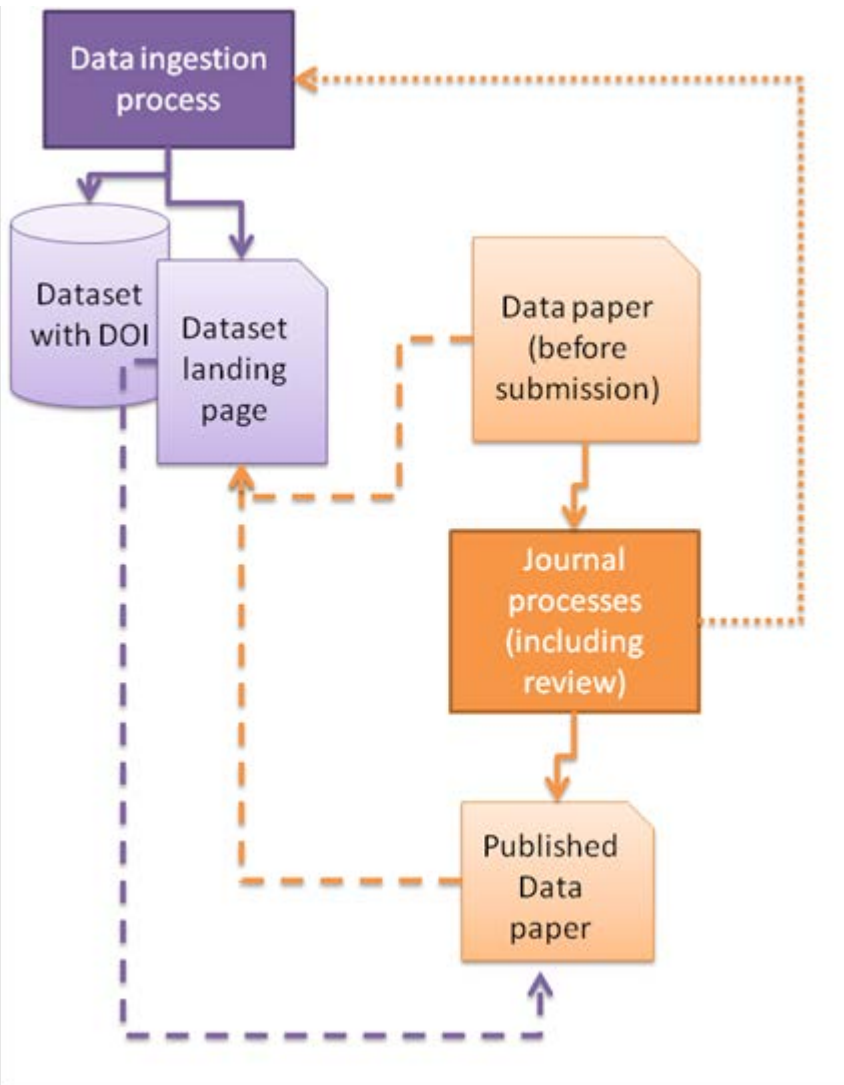
Repository Workflow – NCAR Comp. & Info. Systems Lab Research Data Archive (RDA)

Data Ingest

Check with data provider for changes to files

Notification to provider/user community

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

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.

Embargo

Access Development Phase

Online Data (Most Demanded)

Publish Metadata – User GUIs

Archive (Tape-based)

Not ok

Ok

Distribute metadata

Contact data provider

Errors found

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

Remote backup

GCMD

NCAR CDP

BADC

… OAI-PMH

Geoscience Data Journal workflow

Data set in repository digested and suitable for DOI (passed technical review at data centre)

1 Confirmation and DOI sent to author[1]

2 Author writes Data Paper about the data set, including DOI and submits to GDJ editorial office[2]

Data Paper is reviewed, assessing the following criteria:

1) does it meet the journal's editorial guidelines? Eg data set has a DOI, paper is in scope for the journal

3 If 'no' then Data Paper is rejected[3]

2) scientific review of data set. Eg is it accurate in its methods of data acquisition, statistical info and error calculations etc? Is the data scientifically useful?

If data set does not pass scientific review then Data Paper is rejected. Author should correct/add to the data set and resubmit it to the data centre as a new version. Once the new version of the data set has been ingested the author could submit a new Data Paper (indicating the new data set version number in bibliographic details)[3]

3) review of Data Paper. Eg does the paper adequately describe the data set? Does it explain the data acquisition methodology etc?

Revision required to Data Paper (ie data set ok but not sufficiently described in the paper) → author revises Data Paper and then sends revised version for further review

Data Paper passes review and is accepted for publication in GDJ

Data Paper goes through production at publisher (undergoes copyediting and typesetting, checking of reference details, formatting to journal style, necessary coding and tagging added to allow cross linking, citation and discoverability)

4 XML info (provided by data centre, via author) is used to populate the data set tagging within the article → this appears both fully tagged in the 'data set' section at the start of the article and as a normal reference within the reference list

Data Paper published online in Wiley Online Library (first in Early View, then later within an issue). The Data Paper is assigned its own DOI

5 Citation of the data set by the Data Paper is registered in ISI and other indexing services

Data Paper (and its authors) can accrue citations by other articles citing the data set[5]

6 Data Paper details are sent to the relevant data centre for them to add a cross link from the data set to the Data Paper[4]

Generic data publication workflow.

Dashed lines indicate linking (via URL) or citation (via DOI).

Solid lines indicate the results or inputs into processes.

Dotted line indicated where the results of a process need to be fed back into another process.

Journal responsibilities are orange, data centre's are purple

WP4: Cross-linking
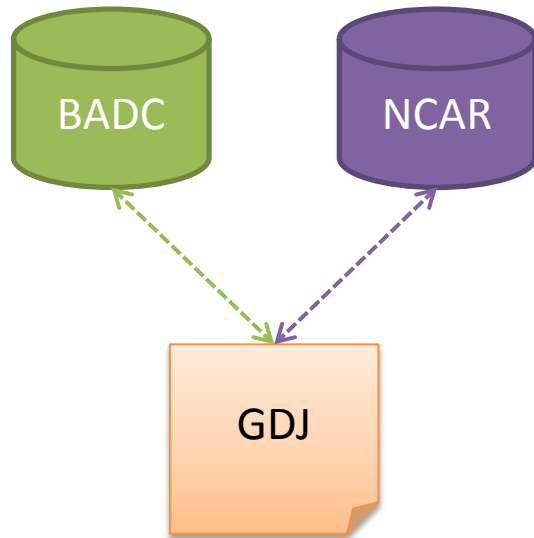
This is what we have to focus on for PREPARDE – demonstrate cross linking between GDJ and a data repository (BADC/NCAR)
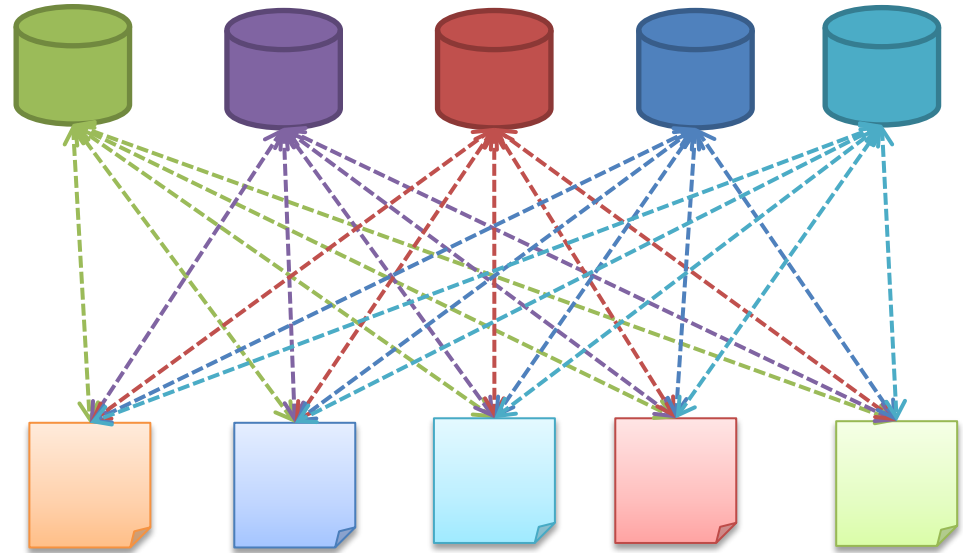
Unfortunately this direct cross-linking isn't scaleable!

Need for off-the shelf solutions that can work across multiple research domains

Cross-linking – the ideal situation
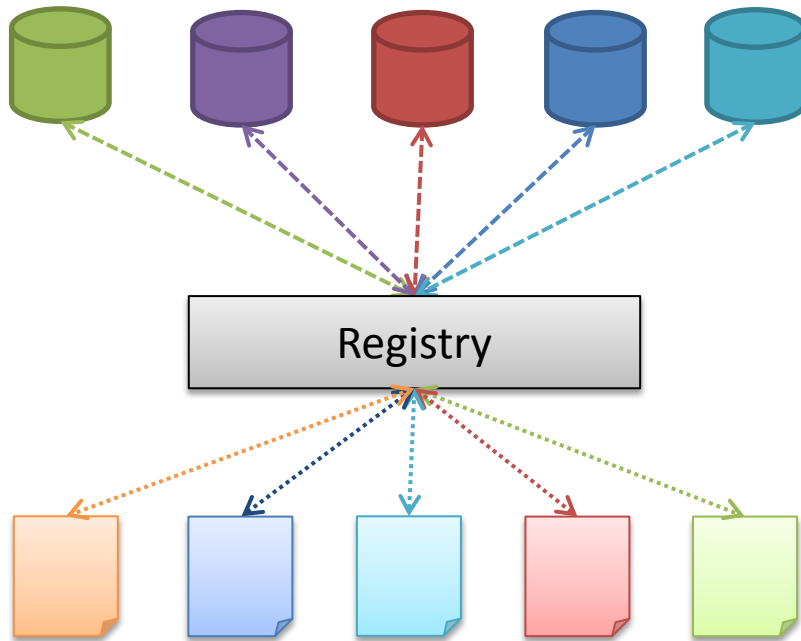
Registry could provide other functions as well as being an intermediary between journals and data repositories like:

- 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? Could DataCite? Could re3data.org?

Registry would need to be discipline agnostic!

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)

Standardised metadata from repositories

DataCite Metadata Store

Journal
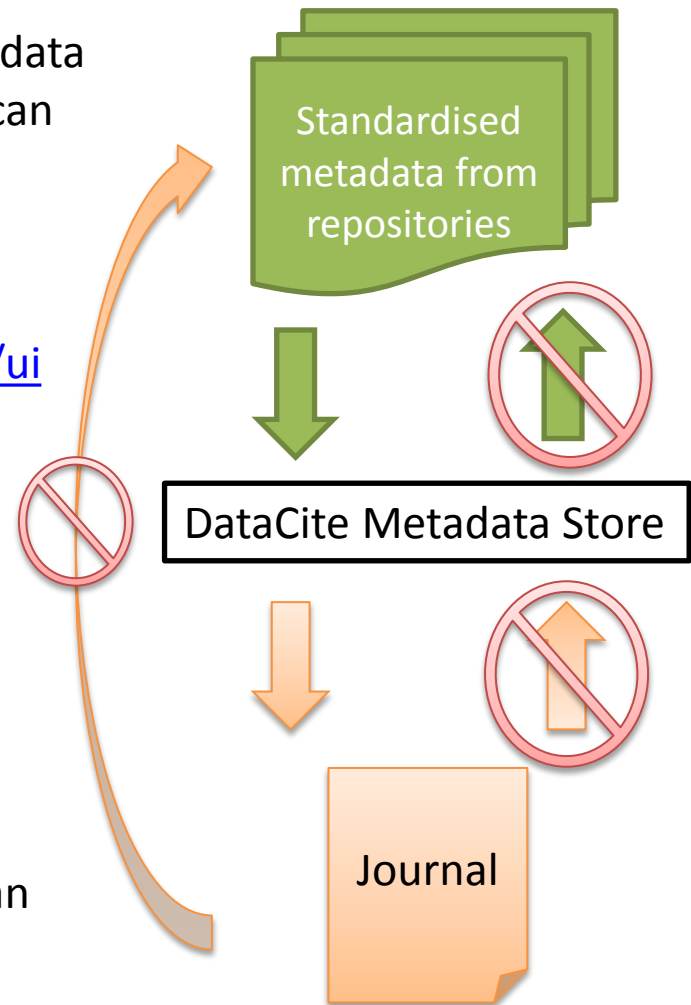
What PREPARDE has done
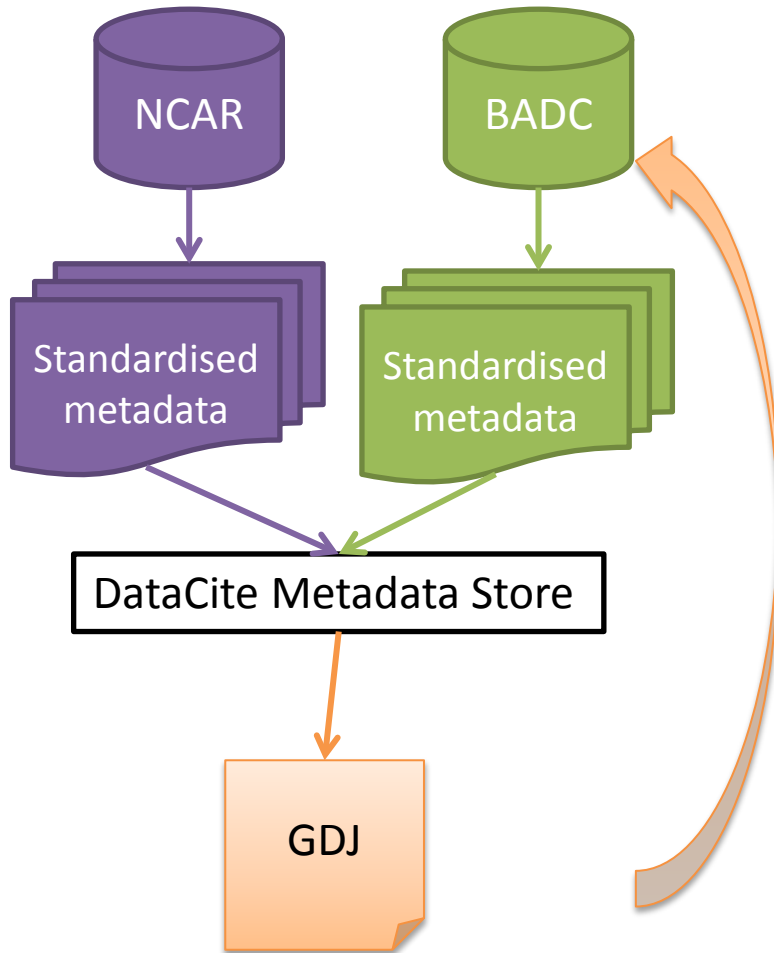
- We already have a link from the GDJ data article to the data repository – thanks to the DOI.

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

- GDJ needs to inform the repository that their dataset has been cited/published – bearing in mind scaling issues!

- At this time, we have a manual work-around (i.e. email)

Live Data paper!

Dataset citation is first thing in the paper (after abstract) and is also included in reference list (to take advantage of citation count systems)

DOI: 10.1002/gdj3.2

WILEY ONLINE LIBRARY

NOT REGISTERED ?
FORGOTTEN PASSWORD ?
INSTITUTIONAL LOGIN >
REMEMBER ME

The Chadwick & RAL Libraries

PUBLICATIONS | BROWSE BY SUBJECT | RESOURCES | ABOUT US

Home > Earth Sciences > General & Introductory Earth Sciences > Geoscience Data Journal > Early View > Abstract

JOURNAL TOOLS
Get New Content Alerts
Get RSS feed
Save to My Profile
Recommend to Your Librarian

JOURNAL MENU
Journal Home

FIND ISSUES

FIND ARTICLES
Early View

FOR CONTRIBUTORS
Author Guidelines
Submit an Article

ABOUT THIS JOURNAL
Society Information
Overview
Editorial Board
Permissions

SPECIAL FEATURES
Data Center FAQs
L F Richardson Award Prize Winners

Open Access License and Copyright
Author FAQs
Article Publication Charges
Wiley Open Access
Institutional and Funder Payments
Guidelines for Reviewers
Guidelines for Repositories

RMetS Royal Meteorological Society

Geoscience Data Journal

Open Access

Data Paper

The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK

S. A. Callaghan[*], J. Waight, J. L. Agnew, C. J. Walden, C. L. Wrench, S. Ventouras

Article first published online: 17 MAR 2013

DOI: 10.1002/gdj3.2

Copyright © 2013 The Authors. Published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes

Issue

Geoscience Data Journal

Early View (Online Version of Record published before inclusion in an issue)

ARTICLE TOOLS
Get PDF (359K)
Save to My Profile
E-mail Link to this Article
Export Citation for this Article
Get Citation Alerts
Request Permissions

Share

SEARCH
In this issue

Advanced >   Saved Searches >

Additional Information (Show All)

How to Cite | Author Information | Publication History | Funding Information

The research presented in this paper was funded by the UK's Ofcom as part of the Spectrum Efficiency Scheme and the support of Ofcom in providing the funding for the GBS experiment is greatly appreciated.

Abstract | **Article** | References | Cited By

Get PDF (359K)

Keywords:
site diversity; radio propagation; fade mitigation techniques

Abstract   Jump to...

The GBS (Global Broadcast Service) dataset is a series of radio attenuation measurements made at three sites in the UK: Chilbolton and Sparsholt, both in southern UK, and Dundee in Scotland. The aim of the experiment was to make long term measurements of the signal strength received from a 20.7 GHz beacon on the US Department of Defense satellite UFO-9 at multiple sites, in order to determine whether the use of site diversity as a fade mitigation technique would be effective. The dataset spans a period of 3 years, from August 2003 to August 2006 with signal attenuation sampled once per second.

Dataset   Jump to...

The GBS (Global Broadcast Service) dataset comes as 3 separate data streams:

- Identifier: doi:10.5285/639A3714-BC74-46A6-9026-64931F355E07

  Creator: Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].

  Title: GBS 20.7 GHz slant path radio propagation measurements, Chilbolton site

  publisher: NERC British Atmospheric Data Centre

  Publication year: 2009

  Resource type: Metadata document

  Version: 1.0

- Identifier: doi:10.5285/db8d8981-1a51-4d6e-81c0-cced9b921390

  Creator: Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].

University of Reading   DCC   University of California CDL California Digital Library   NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

# Dataset catalogue page (and DOI landing page)

Reference to Data Article

Clickable link to Data Article

Problems still to solve

- Automatic methods for:
  - (Data) journal informing repository dataset has been cited
  - Repository linking back to paper citing dataset
- Sharing of dataset metadata between repository and journal
  - So paper author doesn't have to repeatedly enter metadata in multiple locations
  - So corrections made in one place can be propagated across

- Centralised registry for cross-linking
  - Deal with scalability issues in direct linking between journals and repositories
- Methods for issuing corrections to data after data paper has been published
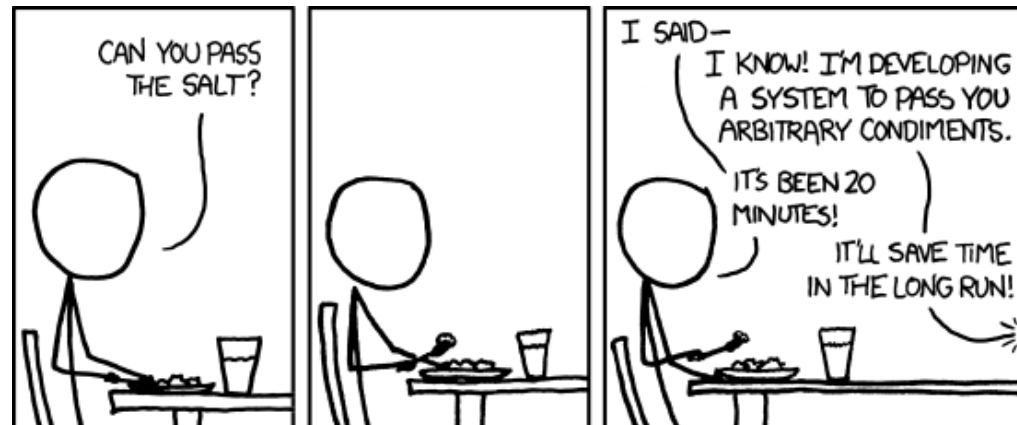
The General Problem http://xkcd.com/974/

Other types of cross-linking

1. Data repository banner ads
2. Geographical maps
3. Pulling metadata from the data repository into journal workflows
4. "Data behind the graph"

Each topic was broken down into the same subsections, which were:
- Type of crosslinking
- Reason for crosslinking
- Current procedures
- How to implement this crosslink in Geoscience Data Journal (GDJ)
- How to roll out this crosslink to other journals
- Further work and issues

Data repository banner ads

Example banner link in a ScienceDirect article
(http://www.sciencedirect.com/science/article/pii/S0921818111001159)

Geographical maps (1)

Example mapping of geolocation metadata in the Pangaea data repository landing page.
(http://doi.pangaea.de/10.1594/PANGAEA.735719)

Geographical maps (2)

Example Elsevier article on ScienceDirect displaying geolocation metadata on a map for the dataset referred to in the article.

Pulling metadata from the data repository into journal workflows

- Least mature method of cross-linking
- Requires significant software development time from both the repository and journal sides.
- Requires many-to-many relationships to be built up to map the dataset metadata appropriately
  - not scalable in the long term
  - third party registry and common standards for dataset metadata could help
- Journal publishers have multiple third party editorial systems in place, so making changes to these systems would be difficult and time consuming.
- How much dataset metadata do reviewers expect to see on the journal site?
  - Less confusing for the reviewers to see dataset metadata on the repository site, rather than mixed in with the article metadata.

"Data Behind the graph" (1)

Example article with interactive viewer for proteins referred to in the article.
(http://www.sciencedirect.com/science/article/pii/S002228361000522X)

"Data Behind the graph" (2)

Example article with table where data is available in a repository.
(http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0066505)

"Data Behind the graph" (3)

Precipitation metrics by site.

	A	B	C	
1	Site*	Annual rainfall (mm)	Proportion Winterrainfall	CV** ann
2	JAV	327	0.42	27.1
3	MDF	326	0.42	26.9
4	WSB	252	0.47	29.7
5	ST	234	0.56	34.8
6	MM	199	0.55	38.3
7	ET	176	0.57	41.0
8	NW	158	0.55	41.1

Table_1

Enlarge Download

3 views   0 shares   cites coming soon

Published on 28 Jun 2013 - 02:42 (GMT)
Filesize is 5.37 KB

This data is part of a published article:
Desert Tortoise (*Gopherus agassizii*) Dietary Specialization Decreases across a Precipitation Gradient

PLOS | ONE   Browse more

Share this: Share 0 Tweet 0 +1 0 Embed*

Cite this: Precipitation metrics by site.
Table_1.xls. Ian W. Murray, Blair O. Wolf. PLOS ONE.
10.1371/journal.pone.0066505.t001. Retrieved 09:10, Aug 22, 2013 (GMT).

*The embed functionality can only be used for non commercial purposes... more

Categories
* Chemistry
* Biological Sciences

Authors
Ian W. Murray
Blair O. Wolf

Description

* WM=Wickenburg Mts.; NW=New Water Mts.; ET=Eagletail Mts.; ST=San Tan Mts.; MM= Maricopa Mts.; WSB=West Silverbell Mts.; MDF=Mother's Day Fire (Rincon Mts.); JAV= Javelina site (Rincon Mts.).

** CV=coefficient of variation.

Comments (0)

You must be logged in to post comments.

Tags
* ecology   * community ecology
* Ecological risk
* Niche construction
* Behavioral ecology
* Global change ecology
* Physiological ecology
* Spatial and landscape ecology
* Terrestrial ecology
* Evolutionary biology   * Zoology
* Herpetology   * Radiochemistry
* isotopes   * metrics

Example of data from previous article's table in a repository.
(http://figshare.com/articles/_P recipitation_metrics_by_site_/7 34897)

University of Reading   DCC   University of California CDL California Digital Library   NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Recommendations from cross-linking WP

There are three main recommendations from this work:

1. Standardisation of metadata
2. Use of DOIs and data citation
3. Role of a centralised, 3rd party registry

WP5 Repository Accreditation

For data publication, a repository must be actively managed in order to:

1. Enable access to the dataset
2. Ensure dataset persistence
3. Ensure dataset stability
4. Enable searching and retrieval of datasets
5. Collect information about repository statistics

Guideline structure

- Guidelines are split into general principles, and subject specific appendices.

- Only the Earth and Life sciences in the appendices at this time

What we learned

- Repository accreditation is a very contentious subject!
  - Repository accreditation schemes exist, but don't havesignificant numbers of members.
  - Reason for the lack of uptake of repository accreditation schemes is not clear.
    - Repositories feel that there is no clear benefit?
    - Accreditation process is unclear or too arduous and/or confusing?

- Repositories seem to be content to rely on their own reputations to demonstrate their suitability as archives for data publication.
  - We think this will change in the near future, as data publication and data stability becomes more important.
  - Further work is needed to identify blockers to the uptake of repository accreditation schemes.

WP1 Project management

Workpackage 1: Project management

Deliverable number	Deliverable title	Status
D1.1	Project Plan (including an Evaluation Plan, QA Plan, Dissemination Plan, and Exit/Sustainability Plan)	Complete
D1.2	Project Web Page on JISC Web Site (including copy of accepted Project Plan)	Complete
D1.3	Project Web Site at Lead Institution	Complete http://www.le.ac.uk/projects/preparde
D1.4	Consortium Agreement (for projects involving more than one institution)	Complete
D1.5	Mid-term report (including financial statement)	merged with final report
D1.6	Final project report (including financial statement)	In draft
D1.7	Completion Report (including financial statement)	In draft

Workpackage 2: Journal and data repository workflows

Deliverable number	Deliverable title	Status
D2.1	Journal workflows from author submission of datasets and papers, through review to publication.	complete
D2.2	Data repository workflows from ingestion of data, through data centre technical review, to DOI assignment to dataset	complete

WORKPACKAGE 3: Scientific review of datasets

Deliverable number	Deliverable title	Status
D3.1	GDJ-specific guidance for reviewers of data submitted to GDJ	ongoing
D3.2	Report and recommendations of suggested best practise for scientific reviewers of datasets	ongoing

WORKPACKAGE 5: Data repository accreditation

Deliverable number	Deliverable title	Status
D5.1	Report on requirements for data centre accreditation.	finalised

WORKPACKAGE 4: Cross-linking between repositories and data publishers

Deliverable number	Deliverable title	Status
D4.1	Roadmap to tighter linking between journal publications and datasets, including data visualisation checks and interface improvements, for review processes and enhanced publications.	Complete, merged with D4.2
D4.2	Worked and operational examples of cross-linking between publications and datasets.	Complete, merged with D4.1
D4.3	Roadmap for implementation of data publication at California Digital Library.	Complete
D4.4	Business plan addressing sustainability of the partnerships and cross-linking between data journals and repositories.	In draft

WORKPACKAGE 6: Stakeholder Engagement and Dissemination

Deliverable number	Deliverable title	Status
D6.1	Workshop with funders, policy makers, researchers, data repositories and other interested parties, discussing the requirements and guidelines for scientific review of data.	Complete - workshop held in Mar 2013 at the British Library. Follow-up workshop held in June 2013
D6.2	Workshop with funders, policy makers, researchers, data repositories and other interested parties, discussing the requirements and guidelines for cross-linking between journal publications and datasets.	Complete - workshop held April 2013. Presentations can be found at CrosslinkingWorkshop
D6.3	Workshop with funders, policy makers, researchers, data repositories and other interested parties, discussing the requirements and guidelines for accreditation of data centres	Complete. Presentations and more details at http://www.dcc.ac.uk/events/idcc13/workshops - scroll down to Workshop 6: Data publishing, peer review and repository accreditation: everyone a winner? The workshop report can also be downloaded from IDCC Data Centre Accreditation workshop report

Financial stuff

- Pretty much all the partners spent their funding completely, and any overspends were covered by matched funding arrangements.

- Leicester have a bit of funding left over – plan to use it to attend RDA meetings in September.