

Connecting data repositories and publishers for data publication

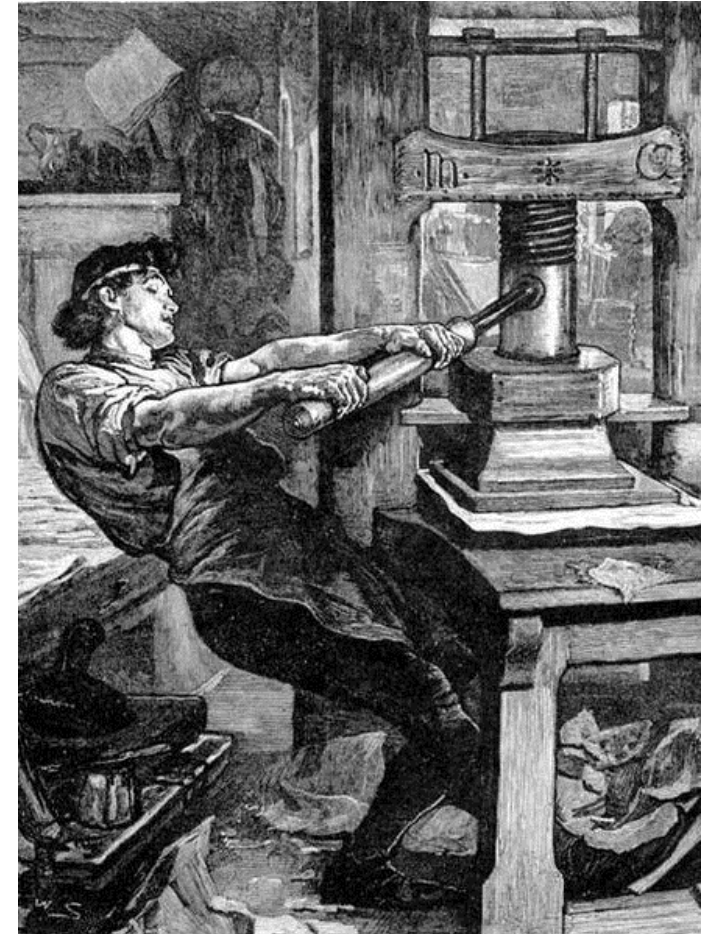
Sarah Callaghan, Fiona Murphy, Jonathan Tedds, Rob Allan, John Kunze, Rebecca Lawrence, Matthew S. Mayernik, Angus Whyte, and the PREPARDE project team

#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

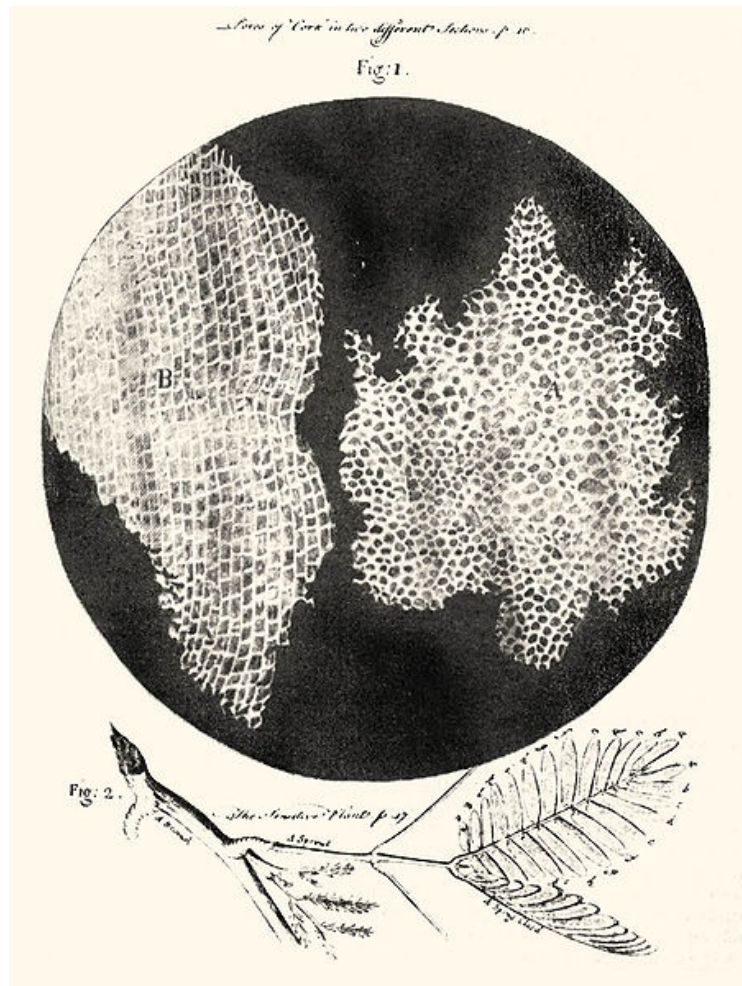


Engraving of printer using the early Gutenberg letter press during the 15th century.

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

[Observations of Stars in the Spiral Nebula. H. 1622.]

The spiral form of this nebula is very distinctly seen in the Pulkova refractor. Unfortunately in the month of March, the best season for the observation of this object, the sky was constantly cloudy; so that I could only get three nights' observations in the months of April and May, when the twilight did not cease for the whole night. It must be attributed to this unfavourable circumstance that the following list of determinations is not so complete as it probably would have been without the twilight. The observations have been made alternately with powers of 138 and 207.

Observations.

Date.	Object.	Magnitude.	Ang. Pos.	No. of measures.	Distance.	No. of measures.
1851, April 7.	N n	14 55'	5	267.1	4
	N a	a = (11)	229 24	3	88.0	3
	N b	b = (11.12)	109 12	3	242.6	3
	a b	93 42	3	298.6	3
April 28.	a b	94 23	3	300.8	4
	N a	228 36	4		
	N b	108 54	4		
	n a	283 42	3		
	n b	153 30	3		
	a d	d = (12.13)	323 51	3		
	N d	277 27	3		
	a e	e = (13)	112 13	3		
	N e	161 56	3		
	N f	f = (12.13)	309 18	3		
	n f	237 31	3		
	a f	335 23	3		
May 3.	a g	g = (12.13)	215 17	3	115.5	4
	a h	h = (12.13)	193 29	3		
	g h	87 5	3		
	N h	h = (12.14)	51 47	3		
	n h	173 29	4		
	b h	317 23	3		
	b l	l = (11.12)	27 20	4		
	n l	83 17	4	355.2	4
	a e	112 56	4		
	N e	161 39	3		
	a m	m = (12.13)	172 43	5		
	N m	190 44	4		
	b m	238 50	4		
	N a	229 12	4	87.0	3
	N n	14 47	4	264.2	3

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”

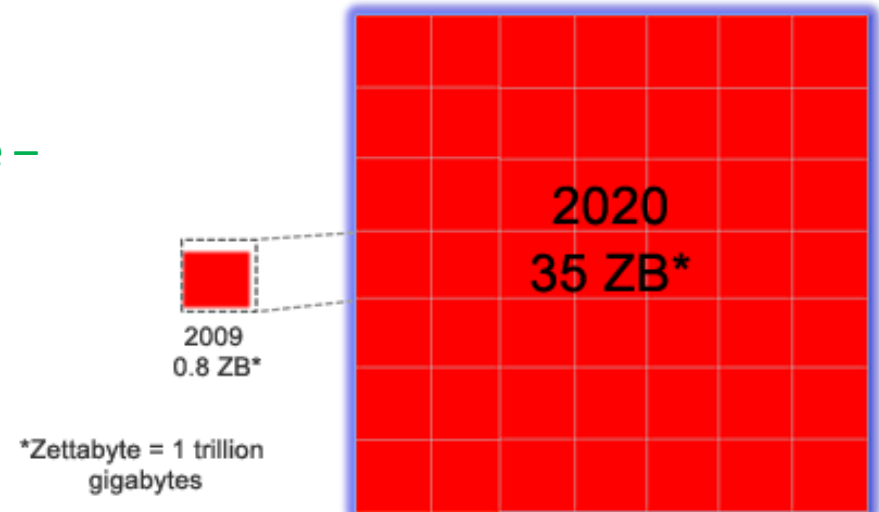


The Digital Universe Decade – Are You Ready?
IDCC White Paper, May 2010

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!

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



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



<http://www.flickr.com/photos/anton41/6588935181/>

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



British Atmospheric
Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL



WILEY-
BLACKWELL



University of
Reading



University of California
CDL
California Digital Library



FACULTY of 1000
POST-PUBLICATION PEER REVIEW



Reasons for citing and publishing data



<http://www.evidencebased-management.com/blog/2011/11/04/new-evidence-on-big-bonuses/>

- 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

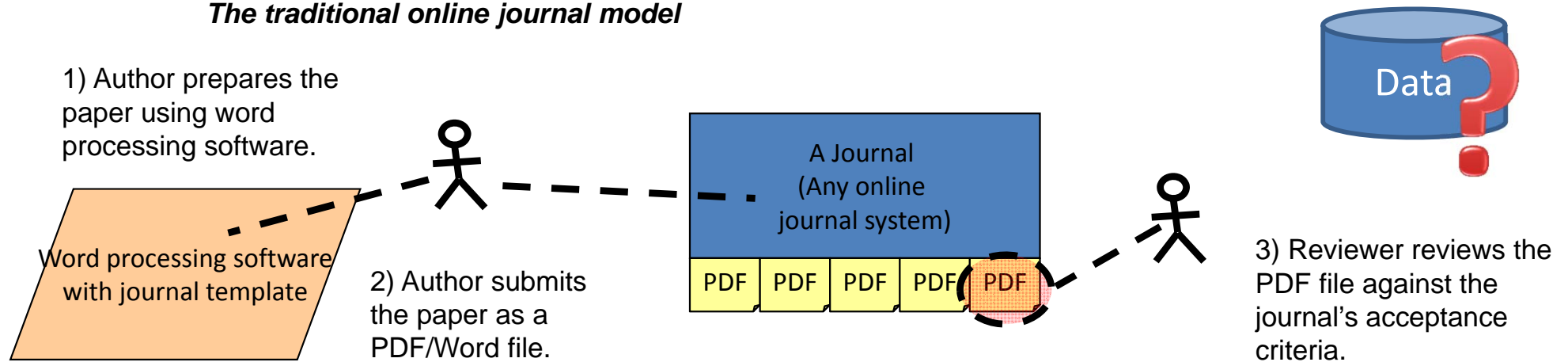


Geoscience Data Journal, Wiley-Blackwell and the Royal Meteorological Society

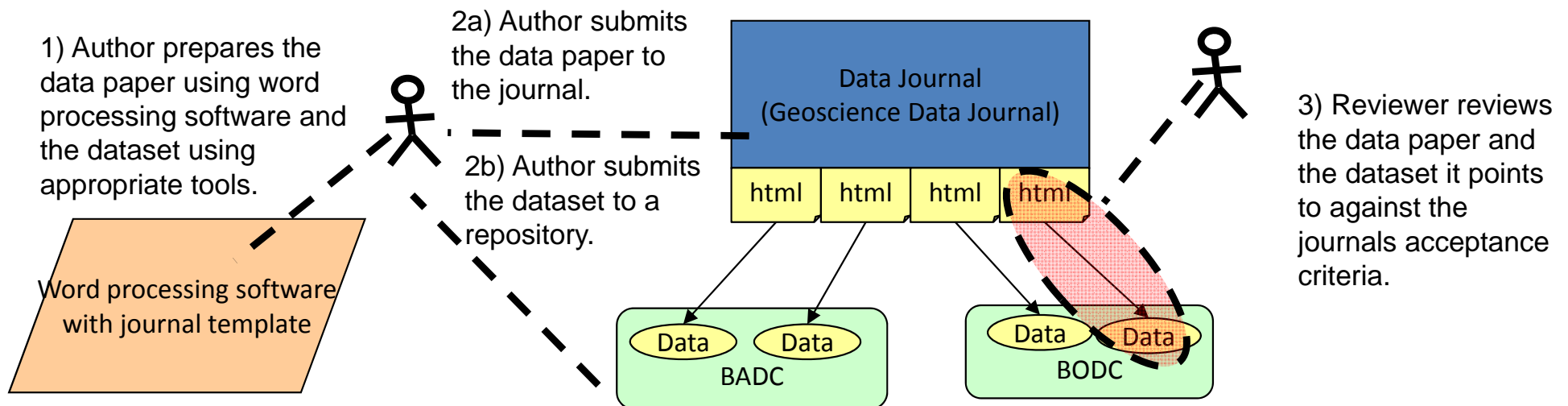
- Partnership formed between **Royal Meteorological Society** and academic publishers **Wiley Blackwell** to develop a mechanism for the formal publication of data in the **Open Access *Geoscience Data Journal***
- GDJ publishes short data articles **cross-linked** to, and **citing**, datasets that have 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 ground breaking conclusions.
 - the **when, how and why** data was collected and what the data-product is.



The traditional online journal model



Overlay journal model for publishing data



JOURNAL TOOLS

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

JOURNAL MENU

Journal Home

FIND ISSUES

Current Issue
All Issues
Virtual Issues

FIND ARTICLES

Early View
Editors' Choice

FOR CONTRIBUTORS

Author Guidelines
Submit an Article

ABOUT THIS JOURNAL

Editorial Board
News
Overview
Permissions
Advertise
Contact

SPECIAL FEATURES

Virtual Issue - Evolutionary Applications to Climate Change
For Reviewers
Cover Gallery
Wiley Job Network
Institutional and Funder Payments
Wiley Open Access
Open Access License and Copyright
Article Publication Charges
Virtual Special Issue - Geoengineering
Virtual Issue - Editor's Choice papers
L F Richardson Award Prize Winners



Data Paper

On the South Atlantic Convergence Zone affecting southern Amazonia in austral summer

Fabien C. Lamaze^{1,*}, Dany Garant², Louis Bernatchez¹

Article first published online: 23 OCT 2012

DOI: 10.1002/asl.401

© 2012 The Authors. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

This is an Open Access article under the terms of the [Creative Commons Attribution Non Commercial License](#) which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Total views since publication: 25 [view chart](#)

Additional Information ([Show All](#))

[How to Cite](#) | [Author Information](#) | [Publication History](#) | [Funding Information](#)

[Abstract](#) | [Article](#) | [References](#) | [Supporting Information](#) | [Cited By](#)

[Get PDF \(504K\)](#)

Keywords:

brook charr; gene expression; hybridization; introgression; quantitative PCR; stocking

Dataset

[Jump to...](#)

GBS (Global Broadcast Service), [doi:10.1029/2007RS003793](#)

Abstract

[Jump to...](#)

Time series analysis of the average rainfall over a target area in southern Amazon Basin showed a spectral peak at 11 day period. An objective method for defining the South Atlantic Convergence Zone (SACZ) is used to identify 28 episodes affecting southern Amazon Basin during the 10 summers in the period 1999–2010. The 28-episode composite precipitation anomalies show significant positive values over the target area. The convergence of moisture over the target area in the SACZ composites is about 35% stronger than the climatological value. Copyright © 2012 Royal Meteorological Society

1. Introduction

[Jump to...](#)

One of the regional scale meteorological systems that affect the weather over a major part of the South American tropics is the South Atlantic Convergence Zone (SACZ). This system somewhat plays the same role for the South American monsoon (Vera *et al.*, 2006; Carvalho *et al.*, 2010) as does the monsoon trough for the South Asian summer monsoon (Keshavamurty and Awade, 1970). The cloud band associated with SACZ extends from the Amazon Basin (Amazonia) to the South Atlantic subtropics, over a stretch of 4000 km or more (Kodama, 1992; Satyamurty *et al.*, 1998) and affects many regions of Brazil with intense rainfall. Some SACZ events are especially intense over interior South American continent (Carvalho *et al.*, 2002, 2004; Muza and Carvalho, 2006), strongly affecting Amazonia.

SEARCH

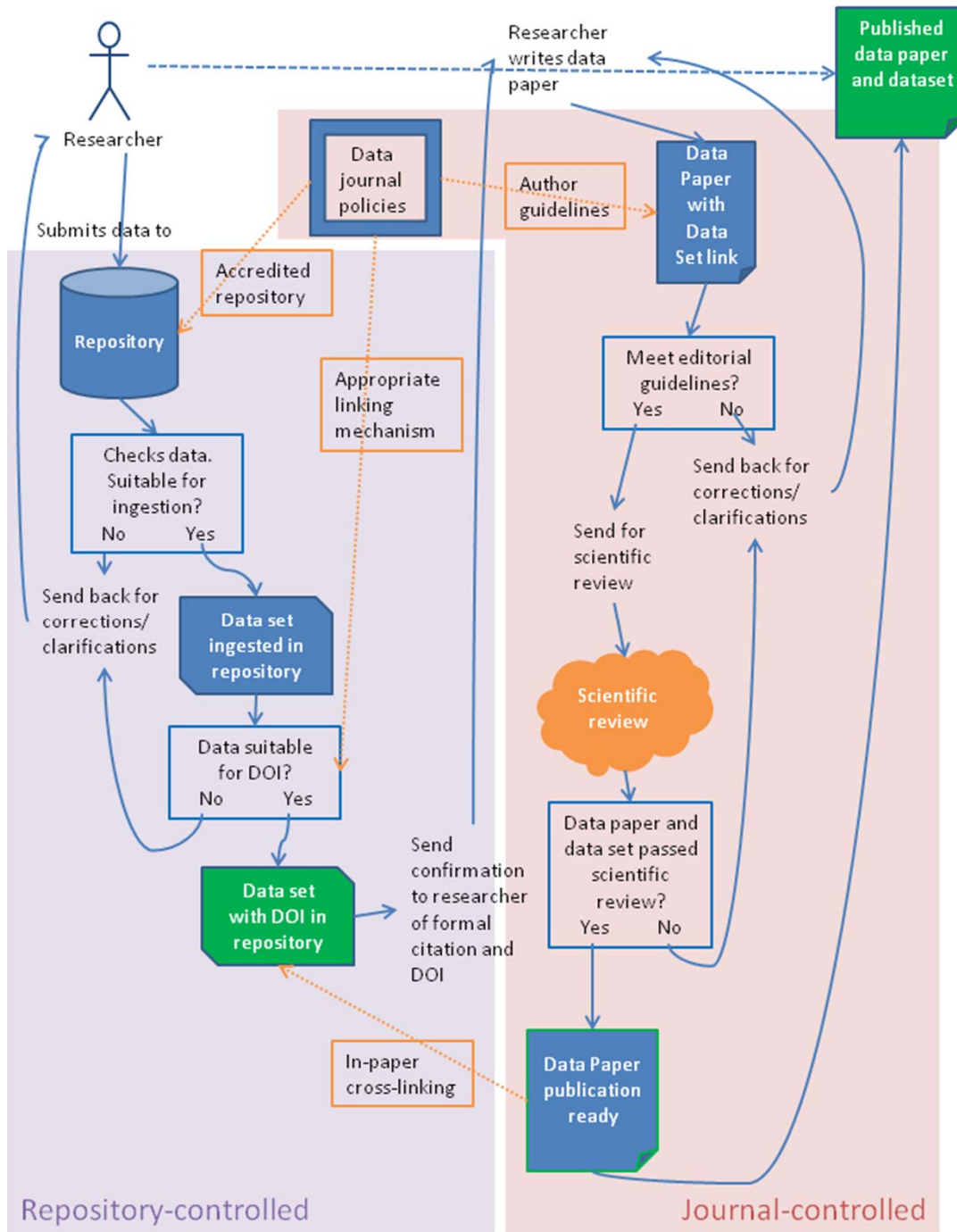
In this issue
Advanced > Saved Searches >

ARTICLE TOOLS

- Get PDF (504K)
- Save to My Profile
- E-mail Link to this Article
- Export Citation for this Article
- Get Citation Alerts
- Request Permissions
- Share

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)



PREPARDE topics

Example steps/workflow required for a researcher to publish a data paper

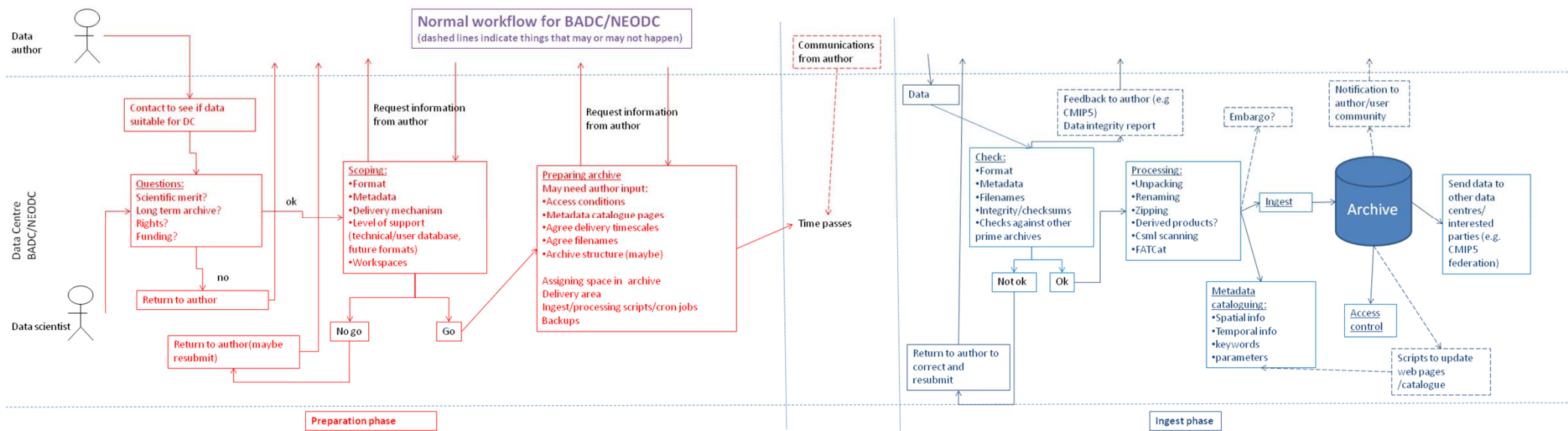
3 main areas of interest (in orange)

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

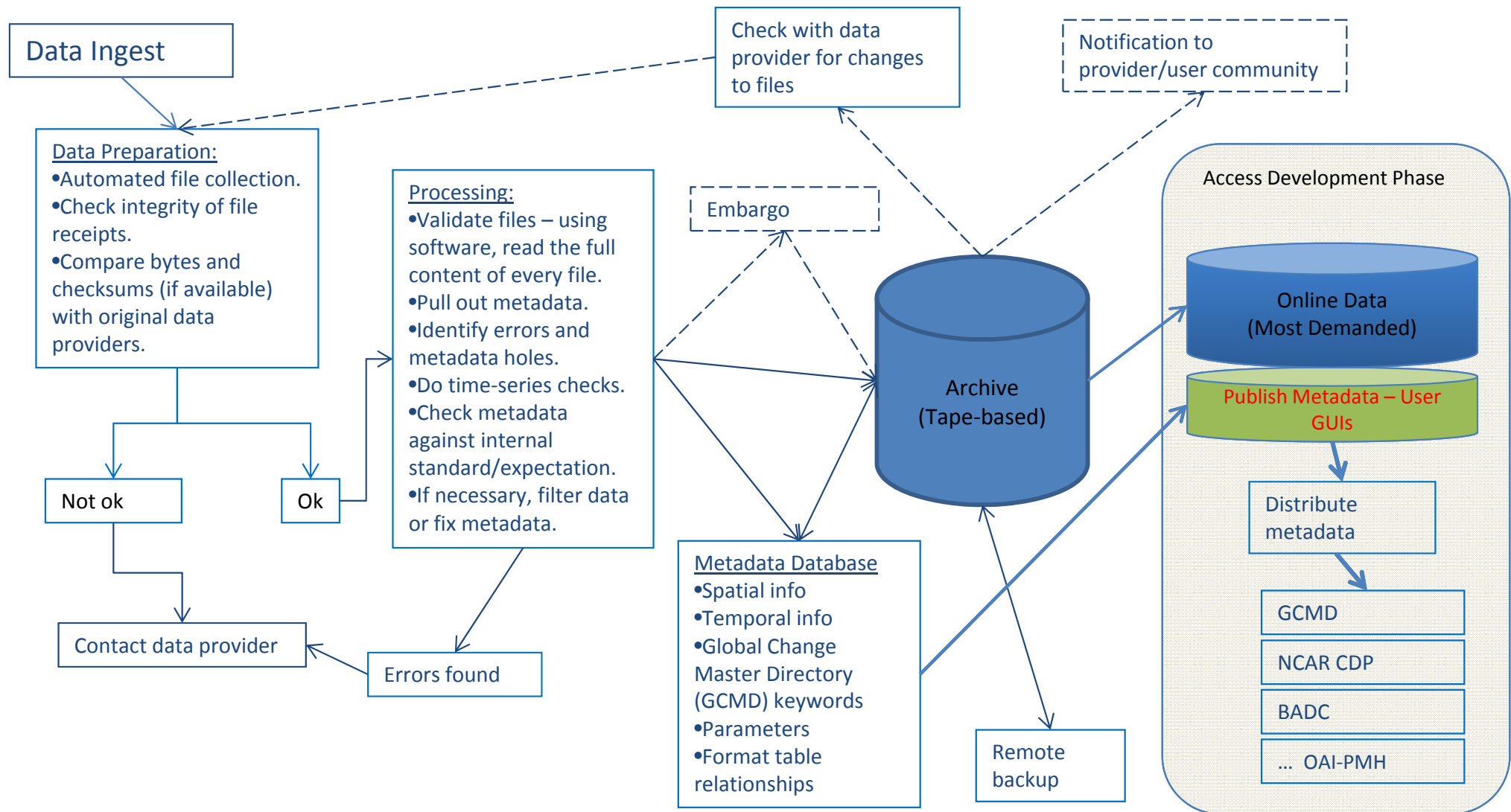
- Division of area of responsibilities between
 - *repository controlled* processes
 - *journal controlled* processes

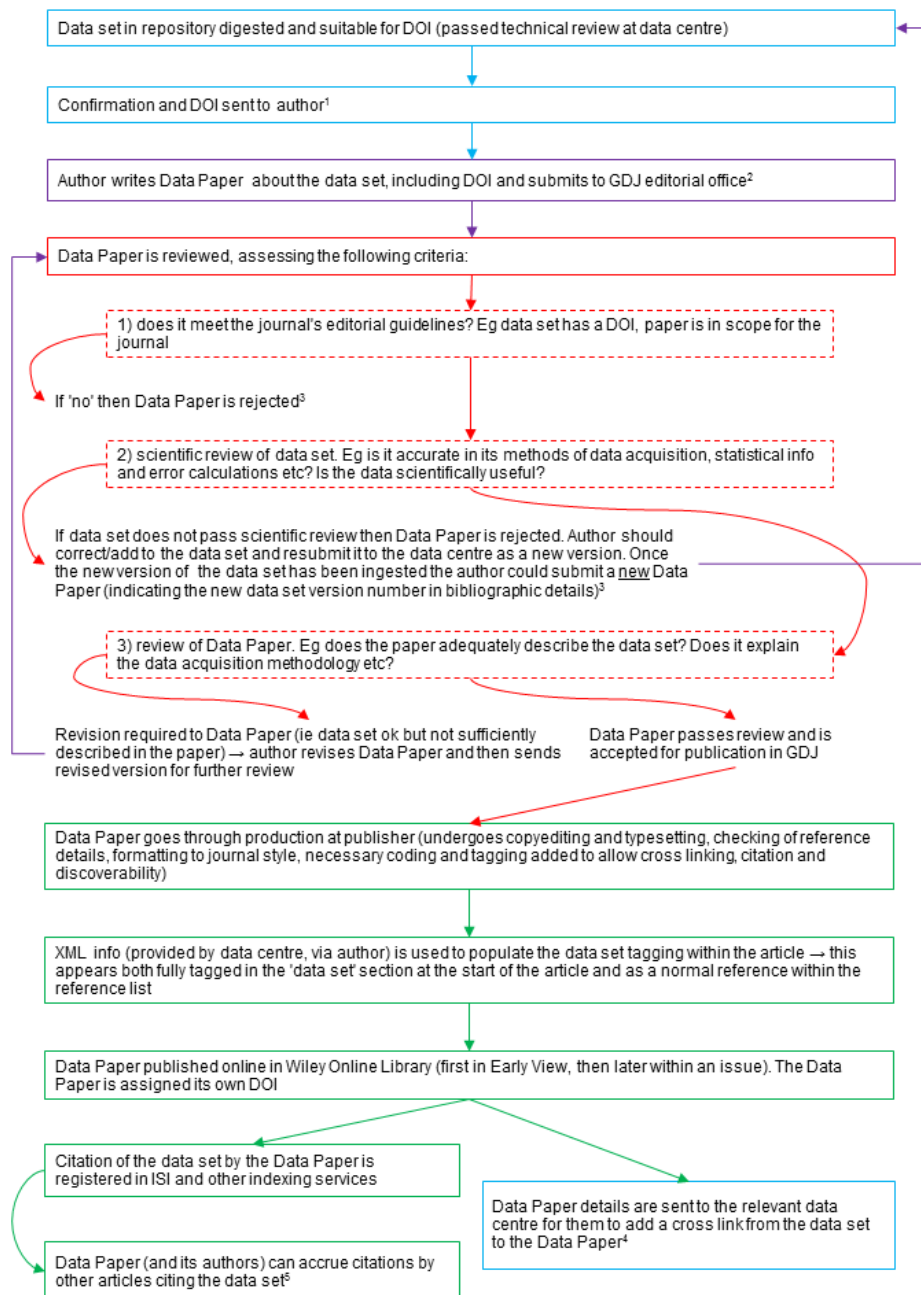
Data repository workflows

- 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)

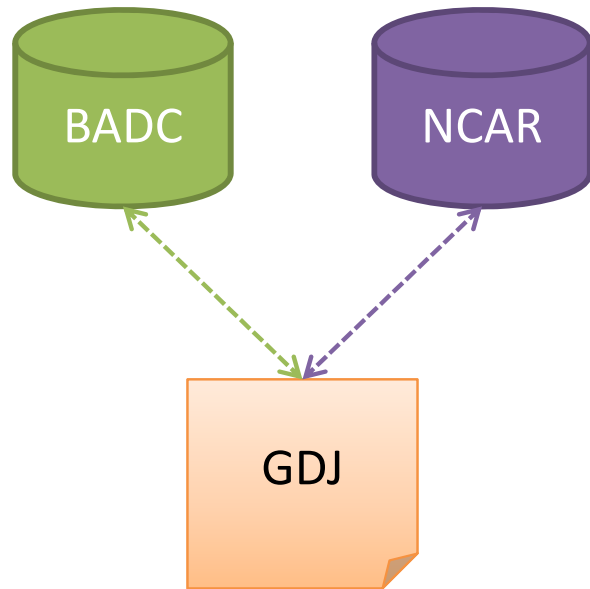




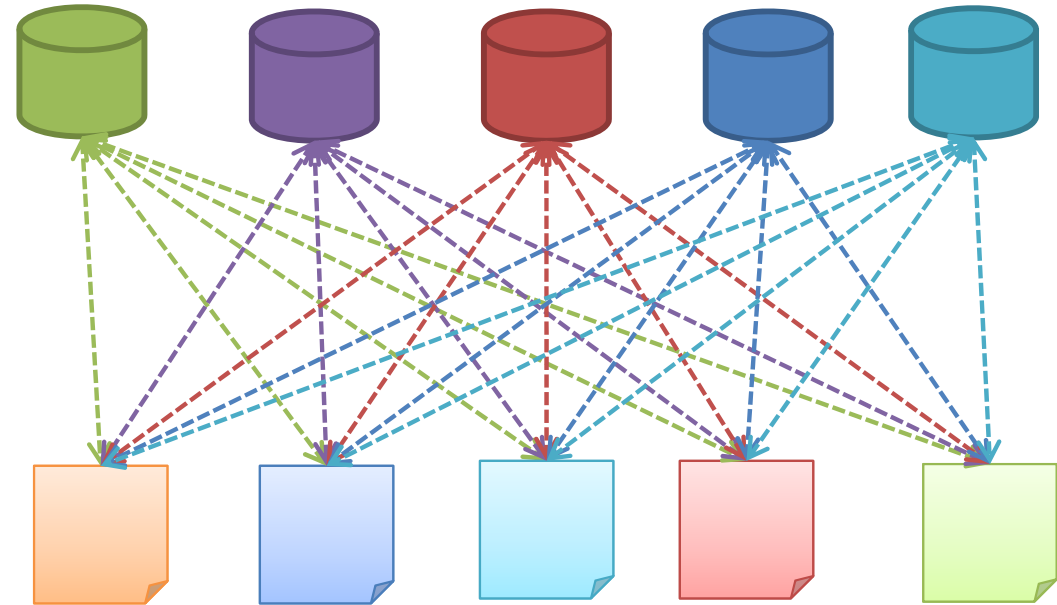
Journal workflow

- 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.

Cross-linking



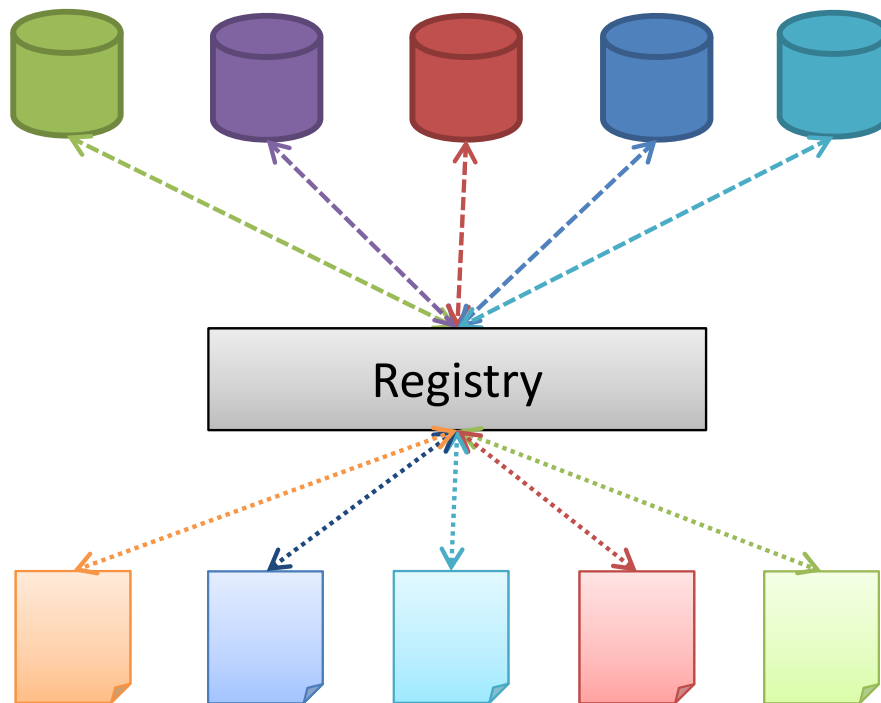
This is what we have to focus on
for PREPARDE – demonstrate
cross linking between GDJ and
BADC (and maybe NCAR)



Unfortunately this direct cross-linking
isn't scalable!

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?

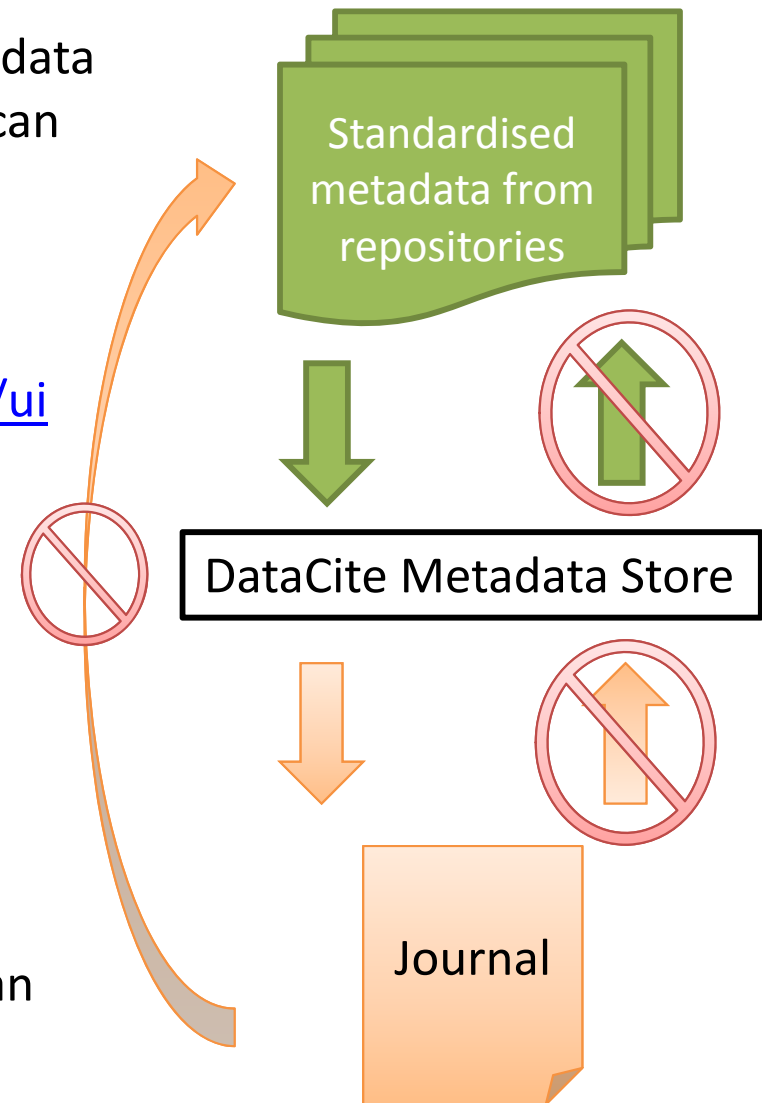
DataCite have standardised a set of bibliometric metadata that have to be submitted before a DOI for a dataset can be minted by a repository.

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

<http://schema.datacite.org/>

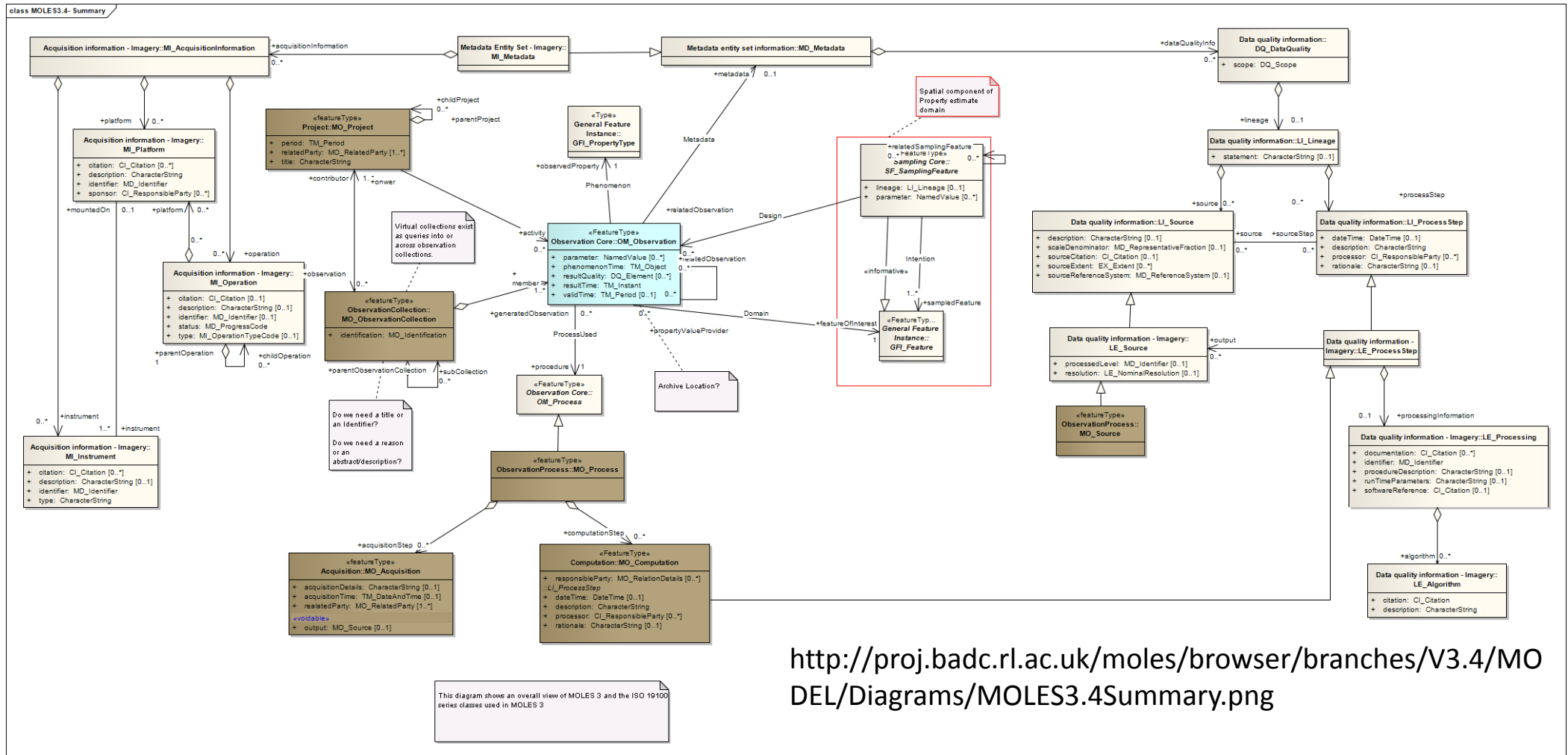
DataCite Mandatory Properties

ID	Property
1	Identifier (with type attribute)
2	Creator (with name identifier attributes)
3	Title (with optional type attribute)
4	Publisher
5	PublicationYear

DataCite Optional Properties

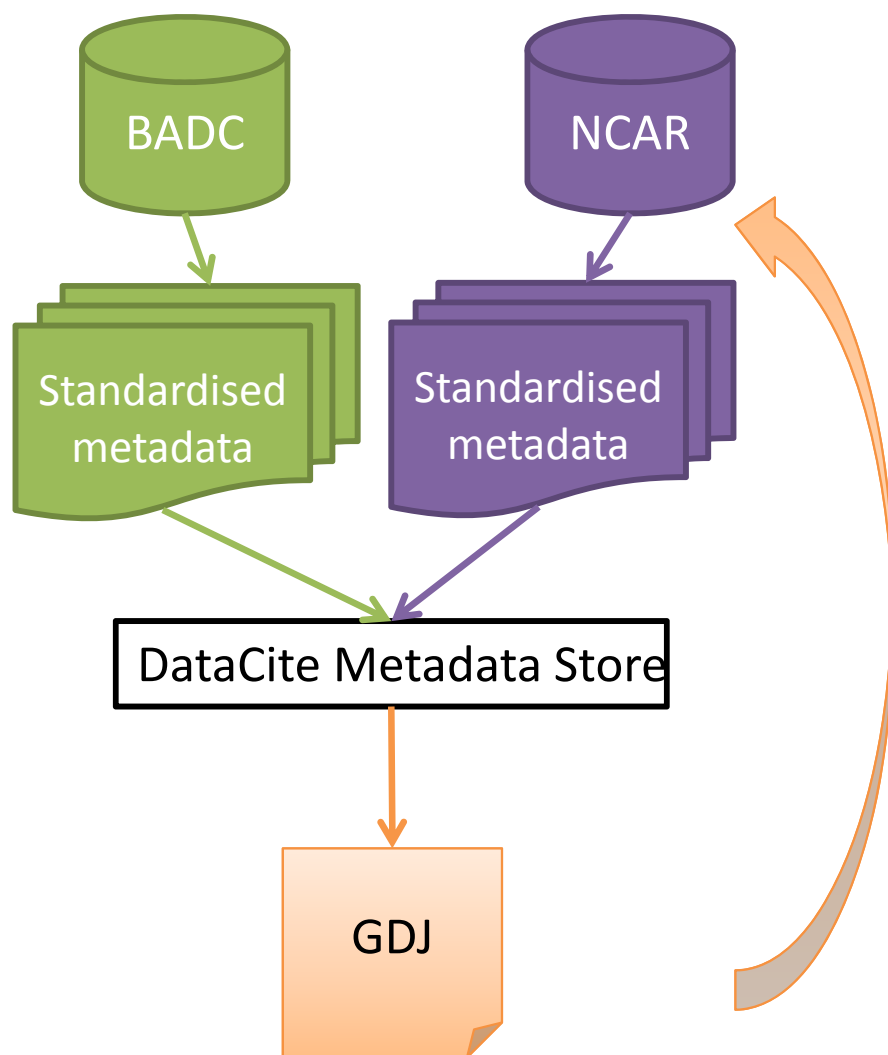
ID	Property
6	Subject (with schema attribute)
7	Contributor (with type and name identifier attributes)
8	Date (with type attribute)
9	Language
10	ResourceType (with description attribute)
11	AlternateIdentifier (with type attribute)
12	RelatedIdentifier (with type and relation type attributes)
13	Size
14	Format
15	Version
16	Rights
17	Description (with type attribute)

MOLES: Metadata Objects for Linking Environmental Sciences v3.4



<http://proj.badc.rl.ac.uk/soles/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!

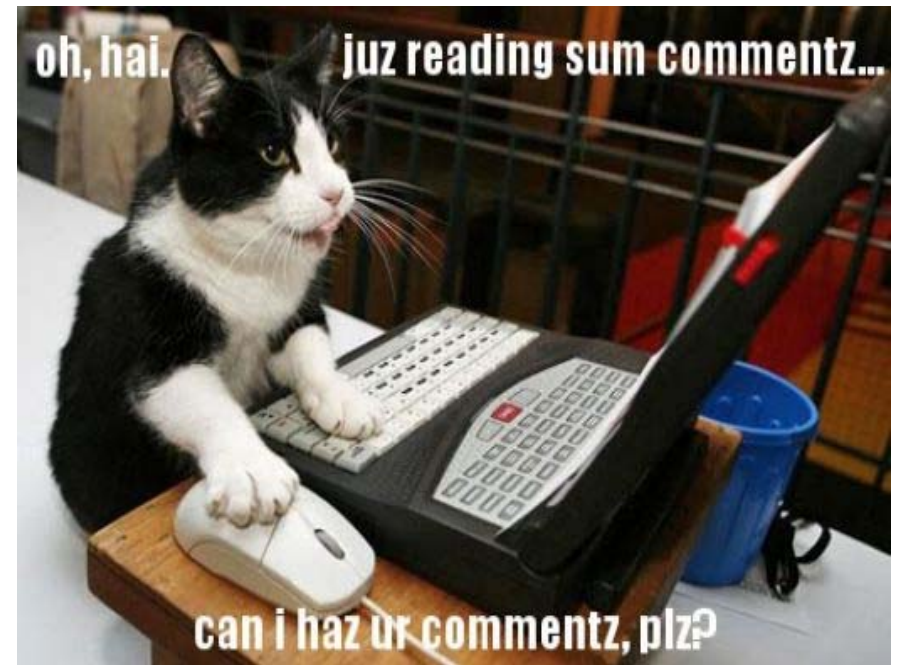


Image Credit: <http://bit.ly/9H4qBX>

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

Project blog: <http://proj.badc.rl.ac.uk/preparde/blog>



British Atmospheric
Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL



WILEY-
BLACKWELL



University of
Reading



University of California
CDL
California Digital Library



Thanks! Any questions?

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

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





University of
Leicester



**British Atmospheric
Data Centre**
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL



**WILEY-
BLACKWELL**



University of
Reading



University of California
CDL
California Digital Library

F1000

FACULTY of 1000
POST-PUBLICATION PEER REVIEW



NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH



Composite image from Flickr user [bnilsen](#) and Matt Stempeck (NOI), shared under [Creative Commons license](#)

