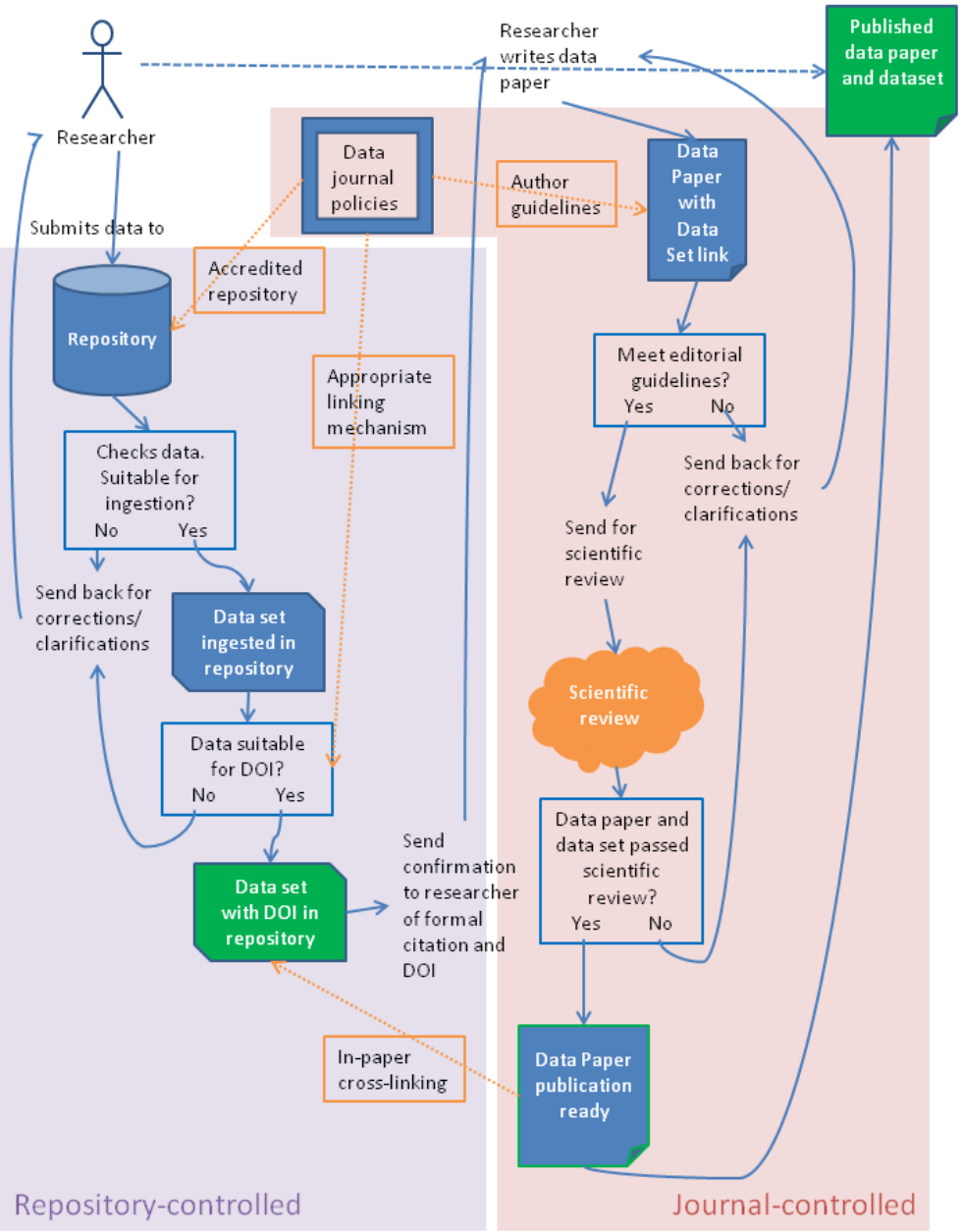
Cross-linking between journal publications and data repositories: a selection of examples

Sarah Callaghan, Jonathan Tedds, Rebecca Lawrence, Fiona Murphy, Timothy Roberts, Will Wilcox and the PREPARDE project team

#preparde

sarah.callaghan@stfc.ac.uk @sorcha_ni

IDCC14, San Francisco, February 2014



PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences

Example 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 http://bit.ly/ZhYHZl
3. Scientific peer-review of data http://bit.ly/DataPRforComment

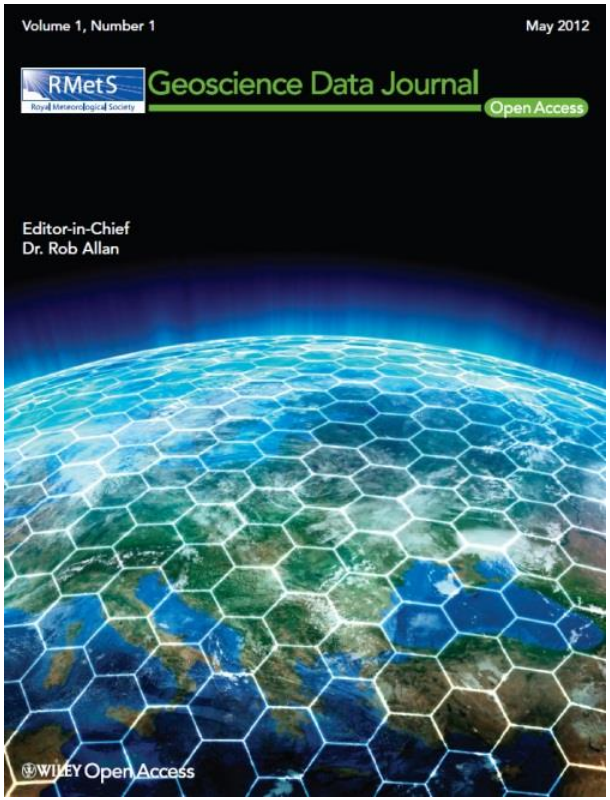
http://proj.badc.rl.ac.uk/preparde/wiki

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.



RMetS
Royal Meteorological Society

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

The GBS dataset: measure

onlinelibrary.wiley.com/doi/10.1002/gdj3.2/full

BADC – Trac | METAFOR | Home | Google Mail | BBC NEWS | News Fr... | Sorcha ní gCeallagh... | Add to Wish List | Other bookmarks

WILEY ONLINE LIBRARY

PUBLICATIONS | BROWSE BY SUBJECT | RESOURCES | ABOUT US

The Chadwick & RAL Libraries

NOT REGISTERED ?
FORGOTTEN PASSWORD ?
INSTITUTIONAL LOGIN >

REMEMBER ME

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)

SEARCH
In this issue
Advanced > Saved Searches >

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

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].

Live Data Article

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

University of Reading | DCC | University of California CDL California Digital Library | NCAR National Center for Atmospheric Research

Dataset catalogue page (and DOI landing page)

Viewing GBS 20.7GHz slan

badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dep_11902119479621181

BADC – Trac | METAFOR | Home | Google Mail | BBC NEWS | News Fr... | Sorcha ní gCeallagh... | Other bookmarks

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Search for in All Go

GBS 20.7GHz slant path radio propagation measurements, Chilbolton site

General Info

Title:	GBS 20.7GHz slant path radio propagation measurements, Chilbolton site
Type:	Activity
Sub-Type:	Deployment
Publication State:	Citable
URI:	http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dep_11902119479621181

Summary

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 20.7GHz 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.

Please cite this dataset as:
Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S. A. Callaghan, J. Waight, C. J. Walden, J. Agnew and S. Ventouras]. GBS 20.7GHz slant path radio propagation measurements, Sparsholt site, [Internet]. British Atmospheric Data Centre, 2003-2005. 1st April 2011. doi:10.5285/639a3714-bc74-46e6-9026-64931f355e07

This dataset is cited in:
S. A. Callaghan, J. Waight, J.L.Agnew, C. J. Walden, C.L.Wrench , S. Ventouras "The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK", Geoscience Data Journal, 17 March 2013, DOI: 10.1002/gdj3.2

Author

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

Online References

Relation	Title
Apply for access	Apply for to GBS data from Chilbolton
Download	Data directory for GBS data from Chilbolton
Documentation	DOI for dataset:10.5285/639a3714-bc74-46e6-9026-64931f355e07
Documentation	Data article in Geoscience Data Journal doi:10.1002/gdj3.2

Associated Data

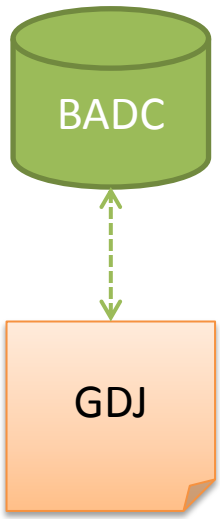
Type	Title
Data Production Tool	Chilbolton: GBS receiver
Activity	Chilbolton Facility for Atmospheric and Radio Research (CFARR)
Observation Station	Chilbolton Facility for Atmospheric and Radio Research (CFARR), UK

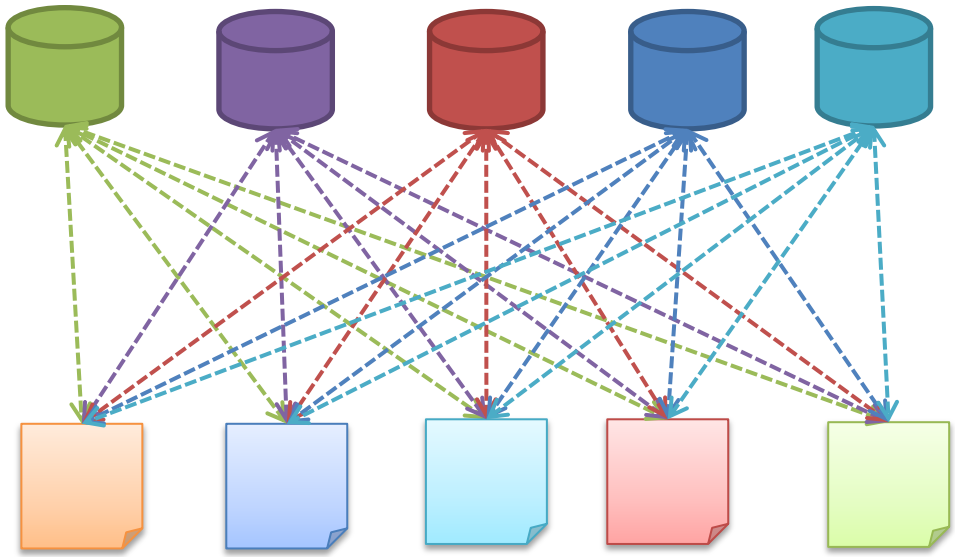
Reference to Data Article

Clickable link to Data Article

JISC | University of Leicester | British Atmospheric Data Centre, NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE, NATURAL ENVIRONMENT RESEARCH COUNCIL | WILEY-BLACKWELL | University of Reading | DCC | University of California CDL California Digital Library | F1000Research | NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Cross-linking





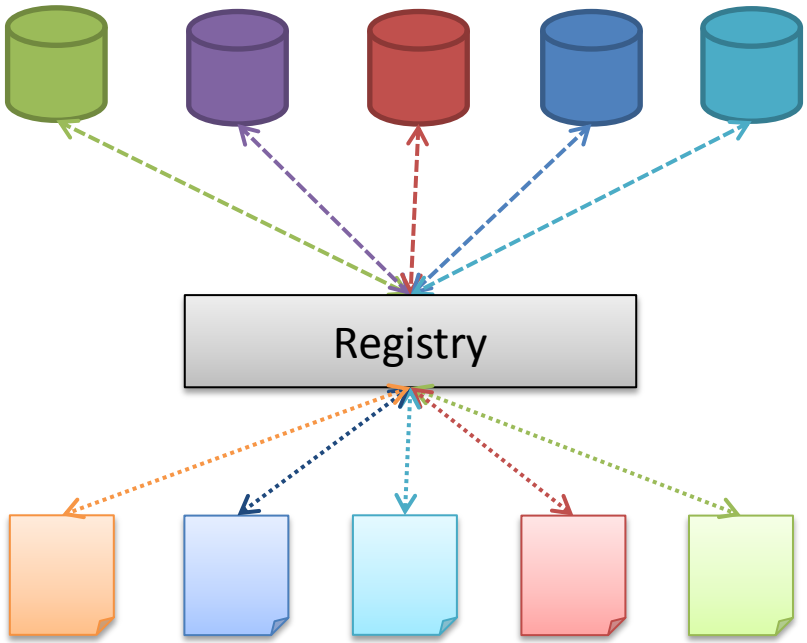
In PREPARDE we demonstrated cross linking between an article in GDJ and a dataset in a data repository (BADC), using data citation via a data DOI

Unfortunately this direct cross-linking isn't scalable!

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

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

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!

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Do we have a start?

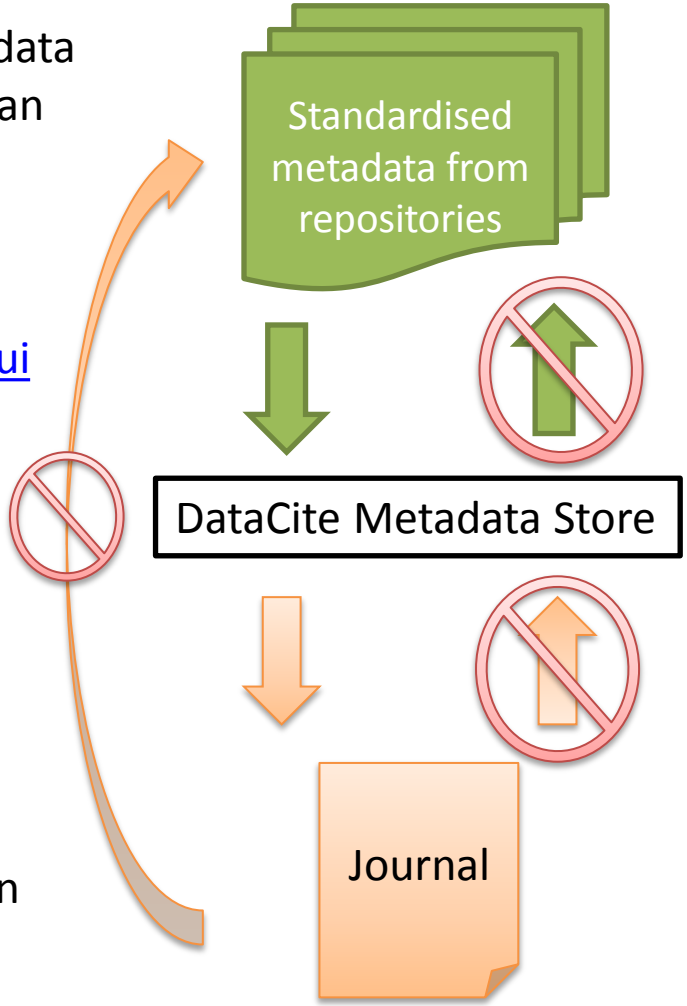
DataCite have standardised a set of bibliometric metadata that have to be submitted before a DOI for a dataset can be minted by a repository.

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

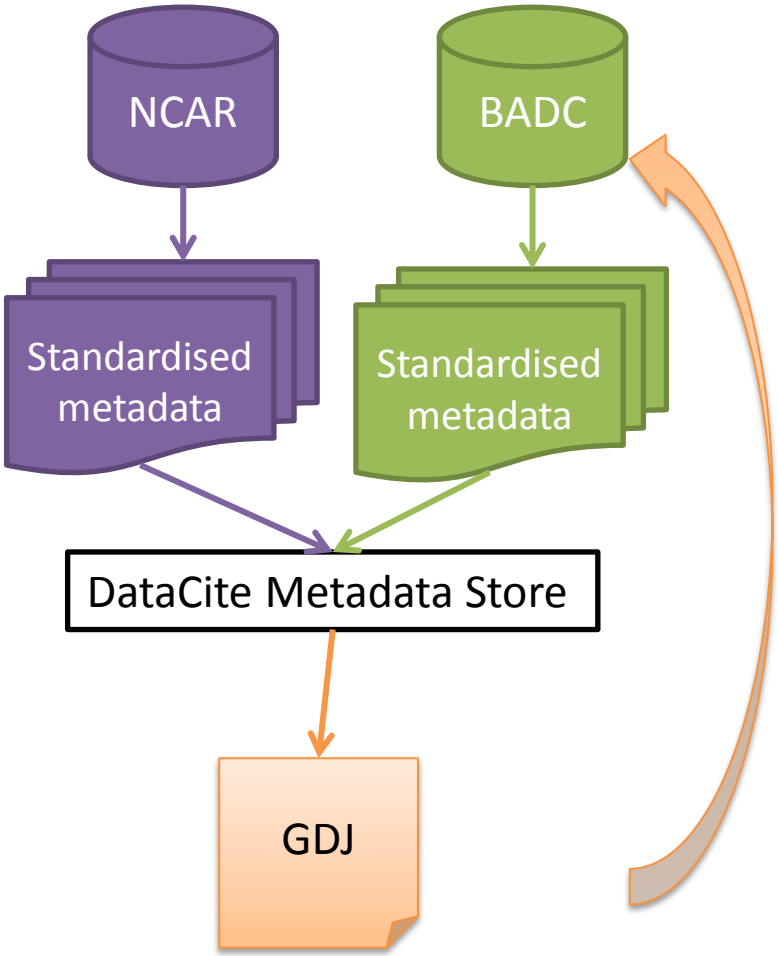
Given a DOI, a journal can then easily find the DOI standard metadata.

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)



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)

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

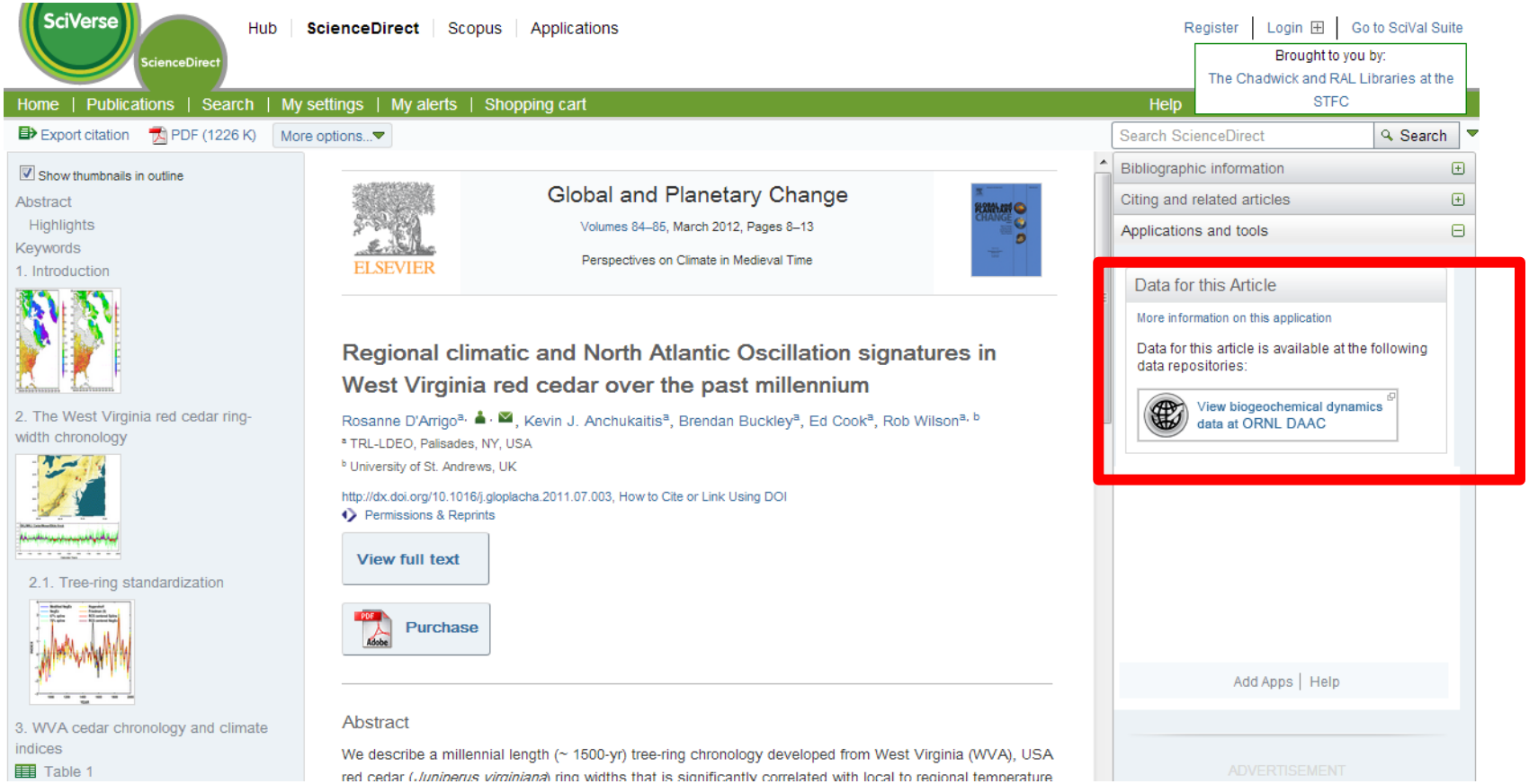
Other types of cross-linking

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

For each type of cross-linking we investigated:

- 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 (1)



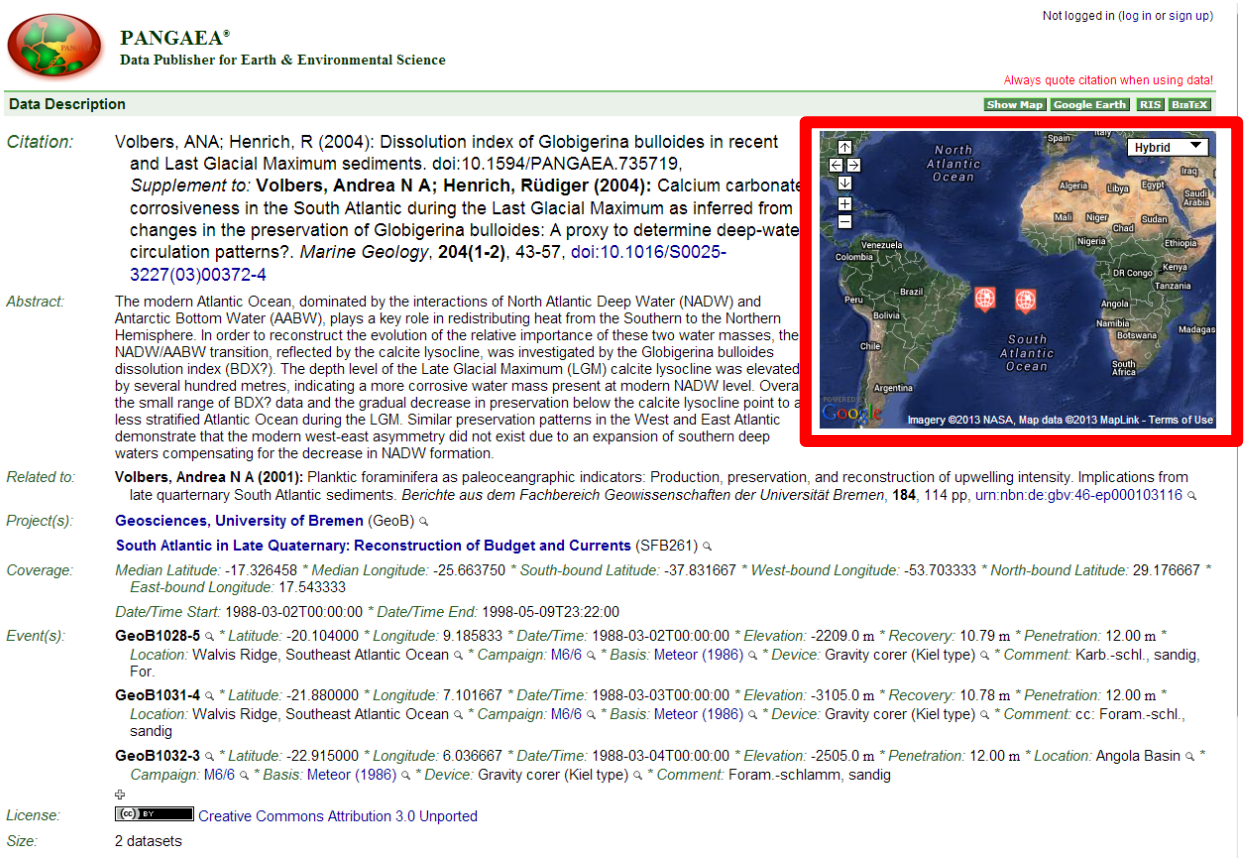
Example banner link in a ScienceDirect article
(http://www.sciencedirect.com/science/article/pii/S0921818111001159)

Data repository banner ads (2)

- Allows readers of the article to get to the dataset in the repository, or to the top level of the repository (where they can browse/search for the data)
- Article is text mined for strings such as flags, accession numbers or names of data repositories
  - A taxonomy and controlled vocabulary will help automate this
- Webpage real estate tends to be limited!
  - Flyover image and link might be more appropriate than fixed ad
- Relationship between journal publishers and repositories to ensure that ad/logo is up to date and link is correct.

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Geographical maps (1)



PANGAEA®
Data Publisher for Earth & Environmental Science

Not logged in (log in or sign up)

Always quote citation when using data!

Data Description

Show Map | Google Earth | RIS | BibTeX

Citation: Volbers, ANA; Henrich, R (2004): Dissolution index of Globigerina bulloides in recent and Last Glacial Maximum sediments. doi:10.1594/PANGAEA.735719, *Supplement to:* **Volbers, Andrea N A; Henrich, Rüdiger (2004):** Calcium carbonate corrosiveness in the South Atlantic during the Last Glacial Maximum as inferred from changes in the preservation of Globigerina bulloides: A proxy to determine deep-wate circulation patterns?. *Marine Geology*, **204(1-2)**, 43-57, doi:10.1016/S0025-3227(03)00372-4

Abstract: The modern Atlantic Ocean, dominated by the interactions of North Atlantic Deep Water (NADW) and Antarctic Bottom Water (AABW), plays a key role in redistributing heat from the Southern to the Northern Hemisphere. In order to reconstruct the evolution of the relative importance of these two water masses, the NADW/AABW transition, reflected by the calcite lysocline, was investigated by the Globigerina bulloides dissolution index (BDX?). The depth level of the Late Glacial Maximum (LGM) calcite lysocline was elevated by several hundred metres, indicating a more corrosive water mass present at modern NADW level. Overa the small range of BDX? data and the gradual decrease in preservation below the calcite lysocline point to a less stratified Atlantic Ocean during the LGM. Similar preservation patterns in the West and East Atlantic demonstrate that the modern west-east asymmetry did not exist due to an expansion of southern deep waters compensating for the decrease in NADW formation.

Related to: **Volbers, Andrea N A (2001):** Planktic foraminifera as paleoceangraphic indicators: Production, preservation, and reconstruction of upwelling intensity. Implications from late quarternary South Atlantic sediments. *Berichte aus dem Fachbereich Geowissenschaften der Universität Bremen*, **184**, 114 pp, urn:nbn:de:gbv:46-ep000103116

Project(s): **Geosciences, University of Bremen** (GeoB)

South Atlantic in Late Quaternary: Reconstruction of Budget and Currents (SFB261)

Coverage: *Median Latitude:* -17.326458 * *Median Longitude:* -25.663750 * *South-bound Latitude:* -37.831667 * *West-bound Longitude:* -53.703333 * *North-bound Latitude:* 29.176667 * *East-bound Longitude:* 17.543333

Date/Time Start: 1988-03-02T00:00:00 * *Date/Time End:* 1998-05-09T23:22:00

Event(s): **GeoB1028-5** * *Latitude:* -20.104000 * *Longitude:* 9.185833 * *Date/Time:* 1988-03-02T00:00:00 * *Elevation:* -2209.0 m * *Recovery:* 10.79 m * *Penetration:* 12.00 m * *Location:* Walvis Ridge, Southeast Atlantic Ocean * *Campaign:* M6/6 * *Basis:* Meteor (1986) * *Device:* Gravity corer (Kiel type) * *Comment:* Karb.-schl., sandig, For.

GeoB1031-4 * *Latitude:* -21.880000 * *Longitude:* 7.101667 * *Date/Time:* 1988-03-03T00:00:00 * *Elevation:* -3105.0 m * *Recovery:* 10.78 m * *Penetration:* 12.00 m * *Location:* Walvis Ridge, Southeast Atlantic Ocean * *Campaign:* M6/6 * *Basis:* Meteor (1986) * *Device:* Gravity corer (Kiel type) * *Comment:* cc: Foram.-schl., sandig

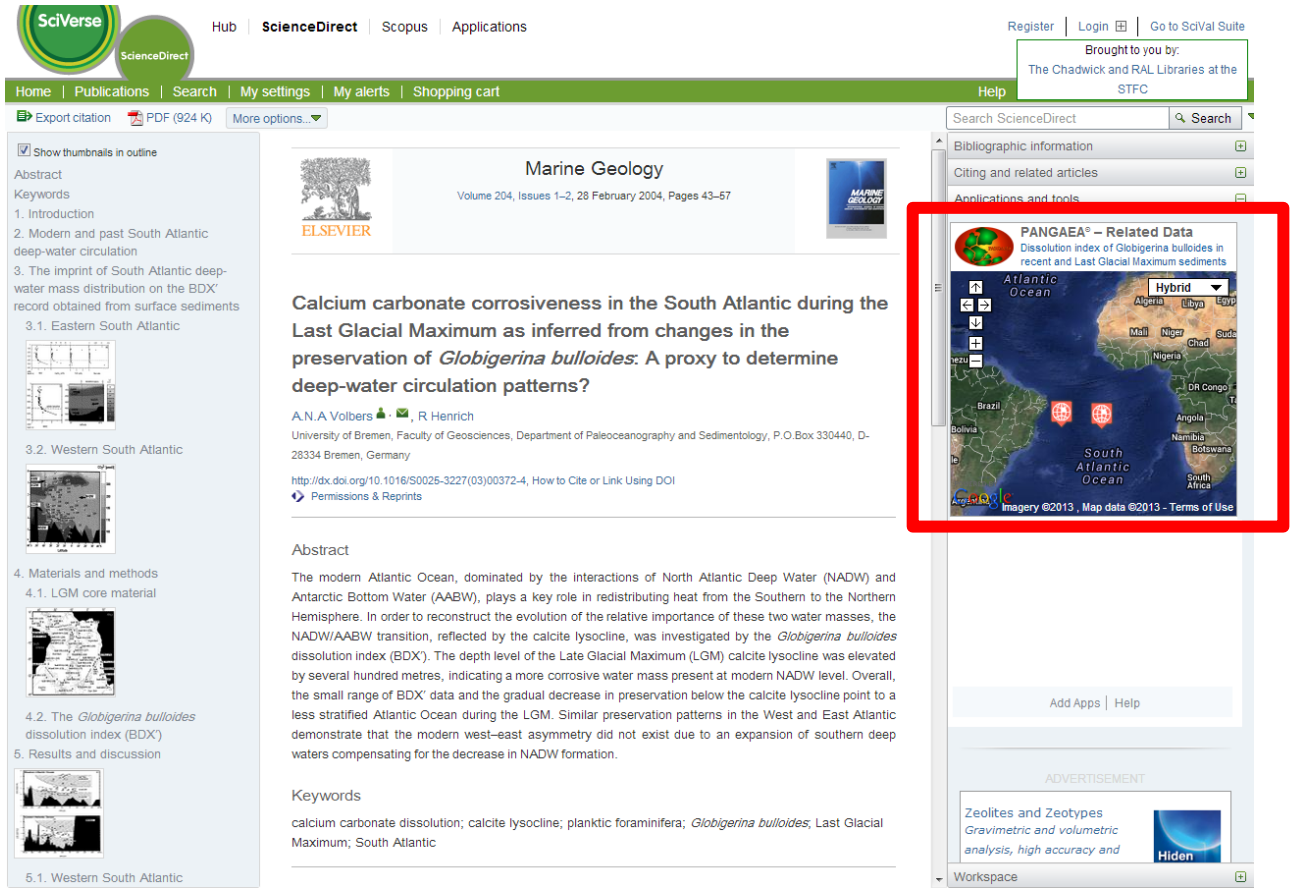
GeoB1032-3 * *Latitude:* -22.915000 * *Longitude:* 6.036667 * *Date/Time:* 1988-03-04T00:00:00 * *Elevation:* -2505.0 m * *Penetration:* 12.00 m * *Location:* Angola Basin * *Campaign:* M6/6 * *Basis:* Meteor (1986) * *Device:* Gravity corer (Kiel type) * *Comment:* Foram.-schlamm, sandig

License: Creative Commons Attribution 3.0 Unported

Size: 2 datasets

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.

Geographical maps (3)

- Takes advantage of geolocation data present in the dataset's metadata
  - Allows the plotting of multiple dataset locations on the same map
- Option to ingest geolocation metadata from repository or from the DataCite metadata
- Best not to duplicate metadata unnecessarily in different locations – i.e. keep metadata about the dataset with the dataset in the repository
- Standardisation is key: ingesting metadata from multiple repositories using different methods for each is not scalable.

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Pulling metadata from the data repository into journal workflows (1)

F1000Research | Article Navigation | Referee Responses 2 | Comments

Extraction of SNPs from exome data

Raw reads were aligned to the reference GRCh37 using bwa 0.61[21]. Local realignment was performed around indels with the Genome Analysis Toolkit (GATK v1.4)[15] framework for variation discovery and genotyping using next-generation DNA sequencing data. Optical and PCR duplicates were marked in BAM files using Picard 1.62[14]. Original HiSeq base quality scores were recalibrated using GATK TableRecalibration and variants called with GATK UnifiedGenotyper. Indels and SNPs were hard-filtered according to Broad Institute best-practice guidelines[22] to eliminate false positive calls and produce the final VCF.

Son exome files
Showing 1/7: Son's Aligned Bam File.bam

1321 views | 3 shares | 96 downloads

```
1  FCB021RACXX:4:1208:7911:79502#CAGATCAT 147 1 11941 0 90M = 11883 -147 CTTCCCGTGTCCTTT
2  FCD044UACXX:4:2205:3896:171755#CAGATCAT 99 1 12059 0 90M = 12212 242 ACTGGAGTGGAGTTTT
3  FCB021RACXX:4:1205:8439:53145#CAGATCAT 99 1 12154 0 90M = 12167 102 ACCACAACCAGGCATAG
4  FCD044UACXX:4:1205:1748:199749#CAGATCAT 163 1 12165 0 90M = 12203 127 GCATAGGGGAAAGAT
5  FCD044UACXX:4:2103:5744:184901#CAGATCAT 99 1 12167 0 90M = 12274 196 GTAGGGGAAAGATTGG
6  FCB021RACXX:4:1205:8439:53145#CAGATCAT 147 1 12167 0 90M = 12154 -102 ATAGGGGAAAGATTGG
7  FCD044UACXX:4:1205:1748:199749#CAGATCAT 83 1 12203 0 90M = 12165 -127 TCAACTTCTCTCACAA
8  FCD044UACXX:4:2205:3896:171755#CAGATCAT 147 1 12212 0 90M = 12059 -242 CTCACAACCTAGGC
9  FCD044UACXX:4:2103:5744:184901#CAGATCAT 147 1 12274 0 90M = 12167 -196 CCCTCGCTCCAGCA
10 FCB021RACXX:4:1107:6633:165696#CAGATCAT 99 1 12275 0 90M = 12313 127 CCTCGCTCCAGCAGC
11 FCB021RACXX:4:1107:6633:165696#CAGATCAT 147 1 12313 0 90M = 12275 -127 GCCATCGGAGCCCA
```

To see the rest of the document click on the icon

figshare 1 / 7 Share Cite Download

The Fastq files represent the raw exome data for the son. The BAM files are derived from the fastq files by aligning the reads using a Burrows-Wheeler Aligner (BWA). The BAM file (.bam) is the binary version of a tab-delimited text file that contains sequence alignment data. The BAM file index (.bai) provides fast random access to the BAM file. The compressed VCF file (.vcf.gz) describes variant calls of the data in text format.

Example figshare widget embedded in an F1000Research paper (http://f1000research.com/articles/1-3/v1).
The widget provides access to the data in figshare, enables the metadata to be previewed within the article, and provides repository metadata about the dataset (namely number of views, shares, downloads etc.)

University of Reading | DCC | University of California CDL California Digital Library | NCAR National Center for Atmospheric Research

Pulling metadata from the data repository into journal workflows (2)

- Pre-publication metadata shared between repositories and journals at article submission stage
  - reduces duplication of effort by author in entering dataset metadata twice
  - ensures consistency of information between journal and repository

Possibilities:

- Author inputs minimal dataset information such as DOI and the journal use the DOI to locate the metadata and add the necessary information into the journal article.
- For repositories requiring significant amounts of metadata – may be possible to create a tool that automatically generates the first draft of a very structured data article.

- Implementation of embedded widgets requires many-to-many relationships to be built up to map the dataset metadata appropriately

- How much dataset metadata should the reviewer see on the journal site?

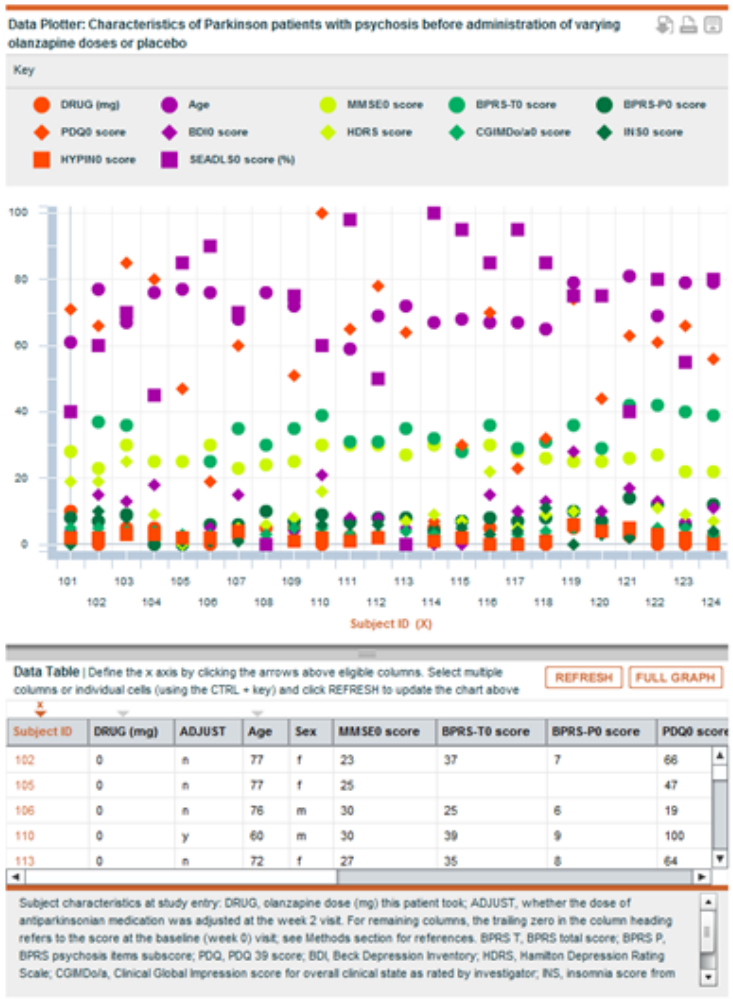
JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

“Data Behind the graph” (1)

	2000	2001	2002	2004	2005	2006	2008	2009	2010	2011	2012	2013	2014
Canada	6.8	7.3	7.7	7.2	6.7	6.3	6.1	8.3	8.0	7.5	7.3	7.2	6.9
France	8.5	7.7	7.9	8.9	8.9	8.8	7.4	9.1	9.3	9.2	9.9	10.7	10.9
Germany	7.5	7.3	8.1	9.8	10.7	9.7	7.2	7.4	6.8	5.8	5.3	5.5	5.6
Italy	10.0	9.0	8.5	8.0	7.7	6.8	6.8	7.8	8.4	8.4	10.6	11.4	11.8
Japan	4.7	5.0	5.4	4.7	4.4	4.1	4.0	5.1	5.1	4.6	4.4	4.4	4.3
United Kingdom	5.5	5.1	5.2	4.8	4.9	5.5	5.7	7.6	7.9	8.1	8.0	8.3	8.0
United States	4.0	4.8	5.8	5.5	5.1	4.6	5.8	9.3	9.6	8.9	8.1	7.8	7.5

Active Chart created and displayed (http://activecharts.org/share/a7dd3bae149b2aba5b8f0d895e00d364), featuring user selection tick boxes to display/hide data and replot it.

“Data Behind the graph” (2)



Left: *F1000Research* data plotting tool showing the raw data and the author graph of that data (Nicholls et al 2013, doi: 10.12688/f1000research.2-150.v1)

Right: Example replot of the data using the *F1000Research* data plotting tool enabling readers to replot raw spreadsheet data, changing the x and y-axis as appropriate

JISC
University of Leicester
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL
WILEY-BLACKWELL
University of Reading
DCC
University of California
CDL
California Digital Library
F1000Research
NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

“Data Behind the graph” (3)

- Accessibility to the data behind the graph enables researchers to make direct comparisons with previously published, or as yet unpublished, results.
- Clicking on the plot would redirect the reader to the subset of the data used to create that plot

- Issues with citation granularity
  - Is it appropriate to assign a DOI to just the subset of a larger dataset that underlies a particular graph?

- Imagine a mixed ecosystem in the future, where repository-managed data, cross-linked with research articles, exists alongside small, specific, image-related datasets that are hosted alongside, and more closely bound to, the articles themselves.

- Relies on authors being willing and able to submit the exact data subsets they used to create each figure
  - Will involve additional work, both in producing the subsets and archiving them

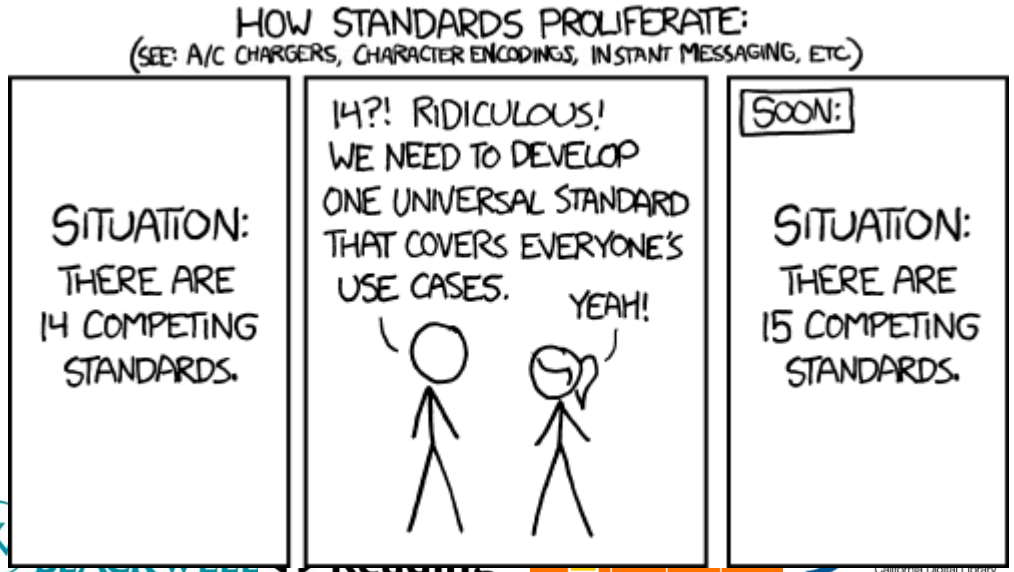
Recommendations and conclusions (1)

Standardisation of metadata

- Automatic processes for the linking and sharing of metadata need to be developed
- Require common standards

PREPARDE recommend the DataCite metadata schema as a common metadata kernel for sharing and exchanging dataset metadata.

It is also recommended that an agreed geolocation standard is implemented given the wide range of multidisciplinary datasets that can be combined in this way.



https://xkcd.com/927/

Recommendations and conclusions (2)

Use of DOIs and data citation

- Use DOIs for linking data to publications
- In the context of a formal data citation

PREPARDE recommends the DataCite citation structure given in the DataCite metadata schema v3.0 http://schema.datacite.org , though where appropriate to the scientific domain, other permanent identifiers may be used.

Citations of data should be included in the references list of the article

Journal's author guidelines should be updated to request authors to cite the datasets used in their article.

Recommendations and conclusions (3)

Role of a centralised, 3rd party registry

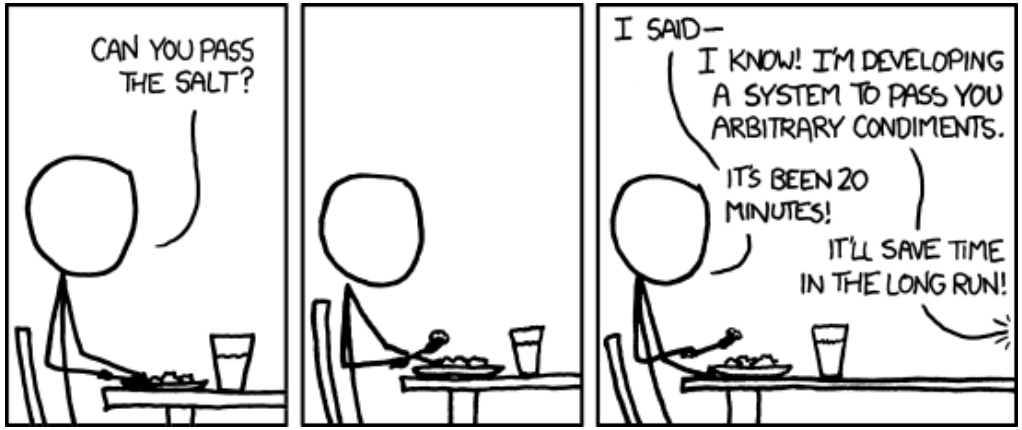
- Simplify the process of passing information between data repositories and journals.

As yet this registry does not exist, though some existing initiatives (DataCite, OpenAIRE) provide some aspects of the service that would be required of this registry.
Although not data-related, CrossRef also provide some aspects of this registry service.

PREPARDE recommend that this be investigated through the Publishing Data Interest Group of the Research Data Alliance

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/

JISC University of Leicester British Atmospheric Data Centre NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE NATURAL ENVIRONMENT RESEARCH COUNCIL WILEY-BLACKWELL University of Reading DCC University of California CDL California Digital Library F1000Research NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH