IJDC | General Article

Cross-Linking Between Journal Publications and Data Repositories: A Selection of Examples

Sarah Callaghan
British Atmospheric Data Centre

Jonathan Tedds
University of Leicester

Rebecca Lawrence
F1000Research

Fiona Murphy
Wiley

Timothy Roberts
Wiley

Will Wilcox
Wiley

Abstract

This article provides a selection of examples of the many ways that a link can be made between a journal article (whether in a data journal or otherwise) and a dataset held in a data repository. In some cases the method of linking is well established, while in others, they have yet to be rolled out uniformly across the journal landscape. We explore ways in which these examples might be implemented in a data journal, such as Geoscience Data Journal, as explored by the PREPARDE project.

Received 01 January 2014 | *Accepted* 26 February 2014

Correspondence should be addressed to Sarah Callaghan, STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot. OX11 0QX, UK. Email: sarah.callaghan@stfc.ac.uk

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v9i1.310
DOI: 10.2218/ijdc.v9i1.310

Introduction

The UK Jisc-funded Peer REview for Publication & Accreditation of Research data in the Earth sciences (PREPARDE) project[1] aimed to investigate the policies and procedures required for the formal publication of research data, in particular focussing on those required for the smooth operation of a data journal (Callaghan et al., 2013). Part of the project investigated the various methods for cross-linking between a data (or other) article and a dataset held in a data repository.

This article discusses cross-linking between journal publishers and data repositories for the purposes of data publication. It identifies a number of possible routes for cross-linking and discusses the issues and barriers associated with them. We discuss the type and reason for the cross-linking approach, current procedures, how to implement the approach, and further work and issues. In all cases, the business case for cross-linking needs to be made; for the paper publishers, if a link takes a great deal of time and effort to make and/or maintain, but only results in a small number of clicks, then that method of cross-linking is not sustainable. In most cases, cross-linking improves visibility of both the dataset catalogue page and the journal article.

Note that this paper is not meant to be a comprehensive survey of all the methods of cross-linking between article and dataset available; rather it is a selection of illustrative examples already in existence. Unfortunately, due to the number and wide variety of types of cross-link and, in some cases, their lack of maturity, it is also not possible to compare and contrast the options to provide definitive answers as to which method is the most appropriate in any given situation. Cross-linking is a part of the wider data publication environment and so is related to other initiatives, including ORCID[2] and Thomson Reuters Data Citation Index[3], though space restrictions limit us from discussing these in any more detail.

Cross-Linking using DOIs (or Other Permanent Identifiers)

For a data journal, permanent linking to the dataset that is the subject of the data article is essential to enable persistent access and appropriate peer review (also investigated in detail by PREPARDE). Data citation also allows datasets linked with any primary scientific article to become part of the formal scientific record, in a transparent way, and allows citation metrics for the datasets to be gathered. These metrics can then provide an indication of the dataset's impact. The citations themselves can be used to track what other uses the dataset is being put to and provide transparency of the research process.

At this time, DOIs in citations are primarily used for journal articles, and the research culture is such that datasets are rarely cited. However, several groups have come together to promote the use of data citation. DataCite is one of these, and acts as a minting authority and registry for the assignment of DOIs to datasets. The British Atmospheric Data Centre (BADC), along with the other NERC environmental data centres and other national, international, institutional and subject-based repositories, are collaborating with DataCite in order to mint DOIs for datasets held in their archives.

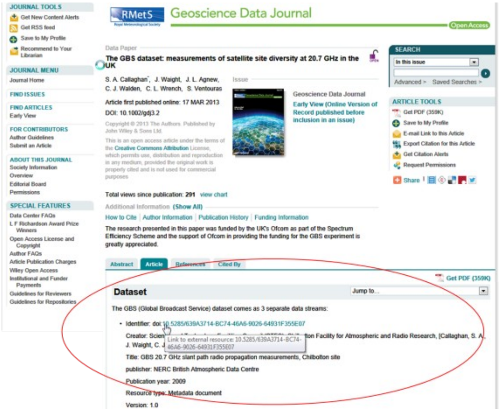
[1] PREPARDE: http://www2.le.ac.uk/projects/preparde

[2] ORCID: http://orcid.org/

[3] Thomson Reuters Data Citation Index: http://wokinfo.com/products_tools/multidisciplinary/dci/

Note that DOIs are not the only permanent identifiers that can be used for data, but they are used as the default in this section because one of the aims of PREPARDE was to demonstrate a working cross-linking system between a data journal (Geoscience Data Journal) and a data repository (BADC), both of which are set up to use DOIs as permanent identifiers for data. A full discussion of alternative identifiers is beyond the scope of this paper, but is very well handled by (Duerr et al., 2011).

Figure 1 shows the main page of a data paper in the Geoscience Data Journal (GDJ). As can be seen, the dataset details are shown at the start of the data paper to make the dataset prominent as the focus of the article. Core elements of the DataCite metadata schema are also displayed so that the details of the dataset are both machine- and human-readable. Figure 2 shows the dataset details in the reference list. The dataset is included as a full reference in the reference list to give it equal weight to other publications, and to allow it to be picked up by citation tracking mechanisms, which only operate on the references list.



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

Geoscience Data Journal

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

Total views since publication: 291 view chart

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.

Dataset

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

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

  Link to external resource: 10.5285/639A3714-BC74-46A6-9026-64931F355E07

  Creator: Scien ton Facility for Atmospheric and Radio Research, [Callaghan, S. A, J. Waight, C. J

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

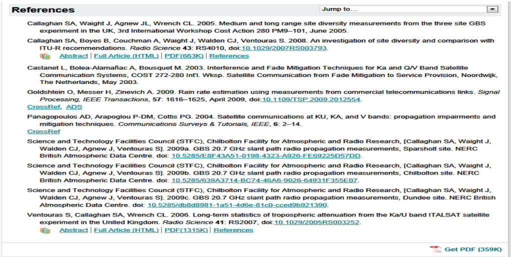
  publisher: NERC British Atmospheric Data Centre

  Publication year: 2009

  Resource type: Metadata document

  Version: 1.0

Figure 1. Screenshot of data article online in GDJ.[4]



References

Callaghan SA, Waight J, Agnew JL, Wrench CL. 2005. Medium and long range site diversity measurements from the three site GBS experiment in the UK, 3rd International Workshop Cost Action 280 PM9–101, June 2005.

Callaghan SA, Boyes B, Couchman A, Waight J, Walden CJ, Ventouras S. 2008. An investigation of site diversity and comparison with ITU-R recommendations. *Radio Science* **43**: RS4010, doi:10.1029/2007RS003793.
Abstract | Full Article (HTML) | PDF(663K) | References

Castanet L, Bolea-Alamañac A, Bousquet M. 2003. Interference and Fade Mitigation Techniques for Ka and Q/V Band Satellite Communication Systems, COST 272-280 Int'l. Wksp. Satellite Communication from Fade Mitigation to Service Provision, Noordwijk, The Netherlands, May 2003.

Goldshtein O, Messer H, Zinevich A. 2009. Rain rate estimation using measurements from commercial telecommunications links. *Signal Processing, IEEE Transactions*, **57**: 1616–1625, April 2009, doi:10.1109/TSP.2009.2012554.
CrossRef, ADS

Panagopoulos AD, Arapoglou P-DM, Cottis PG. 2004. Satellite communications at KU, KA, and V bands: propagation impairments and mitigation techniques. *Communications Surveys & Tutorials, IEEE*, **6**: 2–14.
CrossRef

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan SA, Waight J, Walden CJ, Agnew J, Ventouras S]. 2009a. GBS 20.7 GHz slant path radio propagation measurements, Sparsholt site. NERC British Atmospheric Data Centre. doi: 10.5285/E8F43A51-0198-4323-A926-FE69225D57DD.

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan SA, Waight J, Walden CJ, Agnew J, Ventouras S]. 2009b. GBS 20.7 GHz slant path radio propagation measurements, Chilbolton site. NERC British Atmospheric Data Centre. doi: 10.5285/639A3714-BC74-46A6-9026-64931F355E07.

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan SA, Waight J, Walden CJ, Agnew J, Ventouras S]. 2009c. GBS 20.7 GHz slant path radio propagation measurements, Dundee site. NERC British Atmospheric Data Centre. doi: 10.5285/db8d8981-1a51-4d6e-81c0-cced9b921390.

Ventouras S, Callaghan SA, Wrench CL. 2006. Long-term statistics of tropospheric attenuation from the Ka/U band ITALSAT satellite experiment in the United Kingdom. *Radio Science* **41**: RS2007, doi:10.1029/2005RS003252.
Abstract | Full Article (HTML) | PDF(1315K) | References

Figure 2. Dataset citation in the reference list.[5]

The core metadata elements chosen are also appropriate to the traditional reference structure, e.g. author, publication year, title, publisher. This follows DataCite recommendations for the citation of datasets.

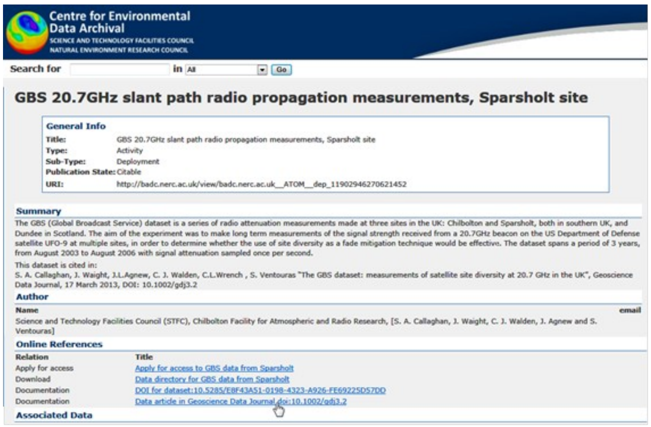
The dataset identifier at the start of the paper is hyperlinked to the dataset landing page using a DOI search. If a DOI is not provided (i.e. an alternative unique identifier, such as an accession code, has been used) then the URL can be hard-coded instead. In the reference list, the reference is linked in the conventional manner – Wiley Online Library (Wiley's online publishing platform) automatically detects the DOI in the reference and uses the DOI resolver service[6] to hyperlink to the cited material (in this case, the dataset). In cases where the dataset does not have a DOI, the link cannot be inserted automatically, but the reference details should be sufficient for the data to be sourced manually. Accession number-based URLs would be hyperlinked, however.

To create the return link from the dataset to the data article, GDJ sends an auto-generated email to inform the data repository (in this example, BADC) when the data article is published (providing the DOI of both the original dataset and of the data article). The data repository then manually (or automatically) updates its dataset landing page with a link to the published data paper (Figure 3).

[4] Taken from: http://onlinelibrary.wiley.com/doi/10.1002/gdj3.2/abstract

[5] Taken from: http://onlinelibrary.wiley.com/doi/10.1002/gdj3.2/references

[6] DOI resolver service: http://dx.doi.org/



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, Sparsholt site

General Info

Title:	GBS 20.7GHz slant path radio propagation measurements, Sparsholt site
Type:	Activity
Sub-Type:	Deployment
Publication State:	Citable
URI:	http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dep_11902946270621452

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.

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 access to GBS data from Sparsholt
Download	Data directory for GBS data from Sparsholt
Documentation	DOI for dataset:10.5285/E8F43A51-0198-4323-A926-FE69225D57DD
Documentation	Data article in Geoscience Data Journal doi:10.1002/gdj3.2

Associated Data

Figure 3. Landing page of the published dataset, showing the citation and link back to the GDJ data article.[7]

The benefit of this cross-linking approach is that it takes advantage of existing mechanisms to turn article DOIs into hyperlinks in the online version of the journal article. As DOIs are functionally identical regardless of what they identify, no new tools or processes need to be created.

The main drawback to this method of cross-linking is that informing the data repository that a dataset held in their archive has been cited via email is not scalable (Figure 4). Hence we propose a registry to act as an intermediary between data centres and journal publishers (Figure 5).

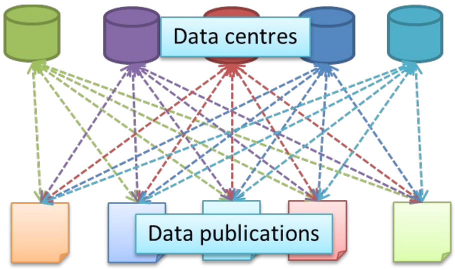


Figure 4. Multiplication of links required for journals and repositories to interact individually.

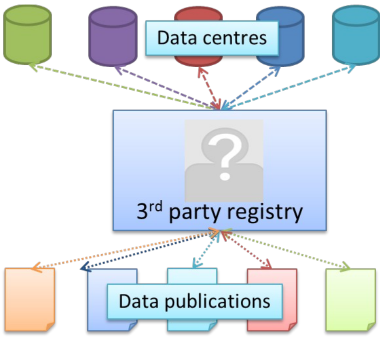


Figure 5. Interactions between data repositories and journals as mediated by a third-party registry.

At this time, no such registry exists, and care will need to be taken to address its sustainability, as it constitutes a single point of failure. This risk is balanced against the benefit of easier and more standardised transfer of information between repository and journal. However, some of the aspects required are already met by the DataCite metadata store. DataCite have standardised a set of bibliometric metadata that they require before a DOI for a dataset can be minted by a repository. These metadata are then made openly available via the DataCite metadata search.[8] This search is also available as an API using Solr Search Handler for the API calls.[9] Given a DOI, a journal

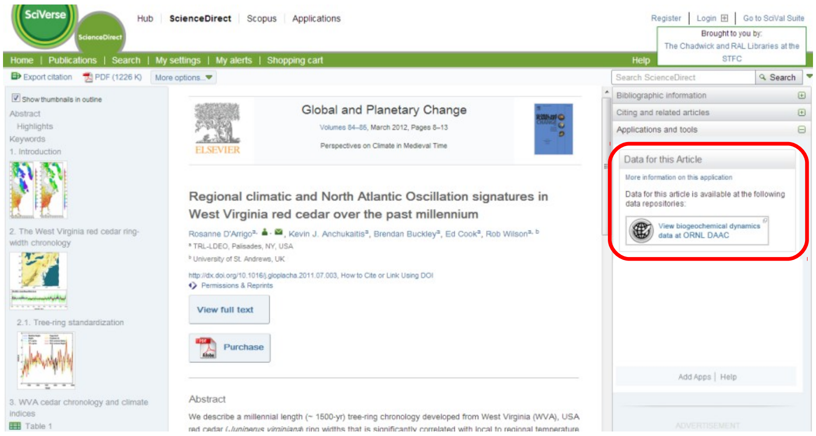
[7] Retrievable from: doi:10.5285/E8F43A51-0198-4323-A926-FE69225D57DD
[8] DataCite metadata search interface: http://search.datacite.org/ui
[9] Endpoint for the DataCite API: http://search.datacite.org/api

can then easily locate the DOI standard metadata. DataCite also has a content resolver.[10] What is missing is the return link, where the journal can inform the repository that a dataset has been cited (directly or via DataCite).The OpenAIRE repository[11] has also been suggested as a potential registry to link between datasets and publications, as they are aiming to collect this linking information as part of their core business[12].

There is a need for research on whether links from dataset records to data articles are followed by users. Journals can be partly achieve this by looking at referrals to identify what proportion come from the dataset landing page. It is worth noting that, in general, links to pages also help with Google page rankings, improving the discoverability of both the article and the dataset. However, Google's current policy means that most regular Google searches will demote journal articles in search results. This statement may therefore only be true of Google Scholar searches.

Data Repository Banner Ads

For articles where data repositories are explicitly mentioned (even if a dataset is not formally cited) a banner ad and link to the data repository could be placed on the article page. This would allow readers of the article to get quickly to the repository hosting the data, where they could search for the data or other information. This has been implemented in some journals, although it is not common practice. For example, Elsevier collaborates with selected data repositories to show banner links next to relevant articles on ScienceDirect (Keall, 2012),[13] providing extra visibility for the data repository (Figure 6).



Brought to you by:
The Chadwick and RAL Libraries at the STFC

Global and Planetary Change
Volumes 84–85, March 2012, Pages 8–13
Perspectives on Climate in Medieval Time

Regional climatic and North Atlantic Oscillation signatures in West Virginia red cedar over the past millennium

Rosanne D'Arrigo[a], Kevin J. Anchukaitis[a], Brendan Buckley[a], Ed Cook[a], Rob Wilson[a, b]

[a] TRL-LDEO, Palisades, NY, USA
[b] University of St. Andrews, UK

http://dx.doi.org/10.1016/j.gloplacha.2011.07.003, How to Cite or Link Using DOI
Permissions & Reprints

Abstract

We describe a millennial length (~ 1500-yr) tree-ring chronology developed from West Virginia (WVA), USA red cedar (*Juniperus virginiana*) ring widths that is significantly correlated with local to regional temperature

Data for this Article
More information on this application
Data for this article is available at the following data repositories:
View biogeochemical dynamics data at ORNL DAAC

Figure 6. Example banner link in a ScienceDirect article (outlined in red).[14]

The data article is text mined for strings such as flags, accession numbers or the names of data repositories. If a string is found to match the name of a repository, then a

[10] DataCite content resolver: http://data.datacite.org/static/index.html

[11] OpenAIRE: http://www.openaire.eu

[12] "Creating a robust, participatory service for the cross-linking of peer-reviewed scientific publications and associated datasets is the principal goal of OpenAIREplus." http://www.openaire.eu/en/component/content/article/76-highlights/326-openaireplus-press-release

[13] ScienceDirect: http://www.elsevier.com/about/content-innovation/database-linking

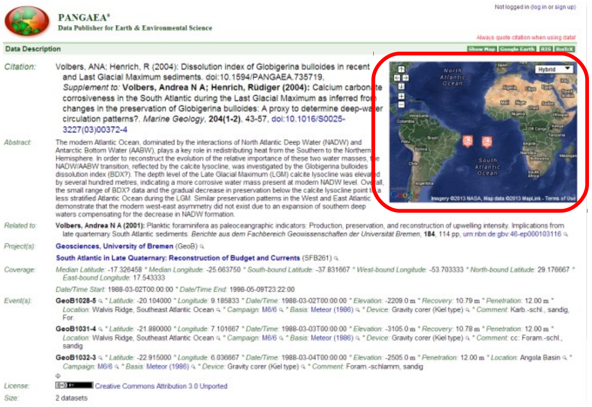
[14] Taken from: http://www.sciencedirect.com/science/article/pii/S0921818111001159

pre-generated banner could be added to the paper sidebar. If the article refers to a dataset using a DOI, then the banner could link directly to the DOI landing page. If not, it should link to the main page of the repository. For efficient text mining, a taxonomy and controlled vocabulary list of repository names and identifiers (such as accession numbers) would need to be created. The main drawback for this method is that webpage real estate tends to be congested, so research would be required to determine whether a fixed ad or a flyover image and link would be more appropriate. In addition, a relationship would be needed between the journal and the repository to ensure that the artwork/logo used for the advert is up to date, and the link used is correct and direct.

Geographical Maps

Another example of cross-linking, particularly relevant to Geoscience Data Journal and other Earth Sciences journals, takes advantage of geolocation data present in the dataset's metadata within the repository (or in the DataCite metadata) to plot the locations on an interactive map, enabling readers to quickly and easily view where the observational data were measured. For papers referring to multiple datasets with geolocation metadata, the positions of the locations can be plotted on the same map, giving an indication of the relative locations.

This form of mapping is currently done by Pangaea (Figure 7) as a standard part of their dataset catalogue and DOI landing pages, using Google Maps as its base layer. Figure 8 shows an example of how that geolocation data from Pangaea is shown within the Elsevier ScienceDirect article webpage.


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
Hybrid
North Atlantic Ocean
South Atlantic Ocean
Imagery ©2013 NASA, Map data ©2013 MapLink - Terms of Use

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-water 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. Overall, 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

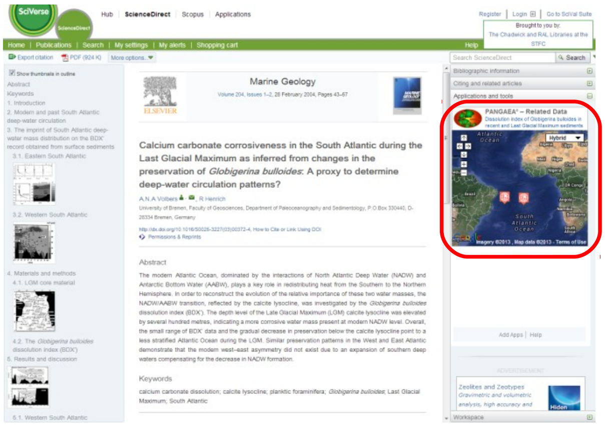
Project(s): **Geosciences, University of Bremen** (GeoB)
South Atlantic in Late Quaternary: Reconstruction of Budget and Currents (SFB261)

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

Figure 7. Example mapping of geolocation metadata in the Pangaea data repository landing page (outlined in red).[15]



Brought to you by:
The Chadwick and RAL Libraries at the STFC

☑ Show thumbnails in outline
Abstract
Keywords
1. Introduction
2. Modern and past South Atlantic deep-water circulation
3. The imprint of South Atlantic deep-water mass distribution on the BDX' record obtained from surface sediments
3.1. Eastern South Atlantic
3.2. Western South Atlantic
4. Materials and methods
4.1. LGM core material
4.2. The *Globigerina bulloides* dissolution index (BDX')
5. Results and discussion
5.1. Western South Atlantic

Marine Geology
Volume 204, Issues 1–2, 28 February 2004, Pages 43–57

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-water circulation patterns?

A.N.A Volbers, R Henrich
University of Bremen, Faculty of Geosciences, Department of Paleoceanography and Sedimentology, P.O.Box 330440, D-28334 Bremen, Germany

http://dx.doi.org/10.1016/S0025-3227(03)00372-4, How to Cite or Link Using DOI
Permissions & Reprints

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. Overall, 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.

Keywords

calcium carbonate dissolution; calcite lysocline; planktic foraminifera; *Globigerina bulloides*; Last Glacial Maximum; South Atlantic

PANGAEA® – Related Data
Dissolution index of Globigerina bulloides in recent and Last Glacial Maximum sediments

Zeolites and Zeotypes
Gravimetric and volumetric analysis, high accuracy and

Figure 8. Example Elsevier article on Science-Direct displaying geolocation metadata on a map (outlined in red) for the dataset referred to in the article.

There are two potential ways of making this cross-link. The first option is to query the dataset's geolocation metadata, as stored in the host repository. Following extraction of this metadata, they can be ingested and added to the dataset metadata, as stored in the journal, and used to plot the dataset spatial extent on a third party map (e.g. Google Maps). The repository would need to provide an API or similar for the journal to query its metadata catalogue. However, because different repositories have different metadata

[15] Taken from: http://doi.pangaea.de/10.1594/PANGAEA.735719

schemas, journals will need to create multiple methods for metadata collection, which is not scalable.

The second option involves a recent development in the DataCite metadata schema (DataCite, 2013) which recommends the use of the GeoLocation property (with point and box sub-properties). This would enable the journal to pull geolocation data directly from the DataCite metadata store along with the other metadata properties (such as title, creator, etc.) that it already ingests, allowing the journal to collect geolocation metadata from multiple repositories without having to map from the different repository metadata schemas to a standard schema. This could potentially allow the spatial information of datasets from different data repositories to be shown on the same map. Clicking on the map could send the reader to the dataset landing page, or allow other features, such as zooming in on an area. Another use case would be to allow journal readers to search for papers based on the geographic location or spatial extent of the datasets referred to in the papers.

We believe it is best for journals to ingest the minimum metadata necessary for citation, and all other metadata to remain in a single source, i.e. the data centre. The data journal would then use an API to look up elements of the metadata that it wanted to make use of, such as geolocation metadata to display a map.

A geolocation cross-link is obviously of use to journals in the geosciences but is also the most obvious way to combine multi-disciplinary datasets and seed new research, as was the case in the Jisc-funded Halogen[16] project. However, the idea of an interactive viewer that could be used for a wide range of data types and fields where an interactive display would be useful to readers.

The main disadvantage of this cross-linking approach is the proliferation of different methods to get the required geolocation data from different repositories. Standardisation is therefore key, and the new Geolocation property in the DataCite metadata schema is a promising first step. The EU’s INSPIRE directive could also be a route for standardisation, although we believe that the DataCite standard may be more accessible and easier to implement.

As with all interactive features, care must be taken to ensure that the value to the journal reader outweighs the effort involved in implementing them.

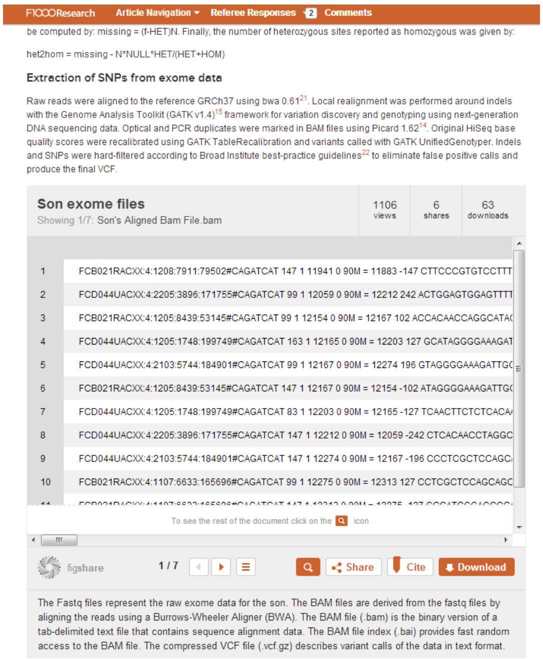
Pulling Metadata from the Data Repository into Journal Workflows

The pre-publication metadata can be shared between repositories and journals at the article submission stage and reduces duplication of effort by the author in entering the dataset metadata twice, and ensures consistency of information between the two sites. For data repositories that require significant quantities of metadata, it may be possible to produce a tool that automatically generates a first draft data article in a highly structured format. Such a tool could even operate to produce a downloadable document suitable for editing in a word processing software ready for uploading (in appropriate format) to the journal submission site. At this time however, the interest in data publication is not such that the effort required to generate this tool is warranted.

An example of sharing metadata between a repository and a journal can be seen in Figure 9, where the figshare widget in the article not only provides access to the dataset used in the article, but also enables the dataset to be previewed using open source

[16] Halogen: http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmp/halogen.aspx

viewers, provides repository metadata about the dataset (namely number of views, shares, downloads etc), and includes its own dataset legend (Lawrence, 2012). Figshare provide F1000Research with a line of xml code to install the widget on the article HTML. This is mainly done by email at the time of writing, but figshare are working on an API for automatic creation of the widget code.



be computed by: missing = (f-HET)N. Finally, the number of heterozygous sites reported as homozygous was given by:

het2hom = missing - N*NULL*HET/(HET+HOM)

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

1106 views | 6 shares | 63 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 ACCACAACCAGGCATA(
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 GTAGGGGAAAGATTG(
6   FCB021RACXX:4:1205:8439:53145#CAGATCAT 147 1 12167 0 90M = 12154 -102 ATAGGGGAAAGATTG(
7   FCD044UACXX:4:1205:1748:199749#CAGATCAT 83 1 12203 0 90M = 12165 -127 TCAACTTCTCTCACA/
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 CCCTCGCTCCAGC/
10  FCB021RACXX:4:1107:6633:165696#CAGATCAT 99 1 12275 0 90M = 12313 127 CCTCGCTCCAGCAGC
```

To see the rest of the document click on the 🔍 icon

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.

Figure 9. Example figshare widget embedded in an F1000Research paper[17]. 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.).

A simple manual workflow could be implemented so that the author inputs minimal dataset information, such as the DOI, and the journal's editorial office or production team could use the DOI to locate the metadata and add the necessary information into the journal article. As a first step, a restricted list of data centres would need to be compiled, and access to this service limited to datasets held in those data centres. The data centre would need to provide an API or other mechanism for the journal to ingest the metadata into the journal submission system (see also the PREPARDE project report on repository accreditation). There would also need to be a standardised mapping between the repository metadata and the journal metadata, e.g. core elements of the DataCite metadata schema.

This method of cross-linking is currently being investigated with a view to implementation by several other generic repositories, including Dryad, Zenodo and DataVerse. Again, implementation of these embedded widgets requires many-to-many relationships to be built up to map the dataset metadata appropriately, which is not scalable in the long term, though a third party registry and common standards for dataset metadata could go a long way to alleviate this. Standards would also allow automatic ingestion and mapping of metadata. Journal publishers often have multiple editorial systems in place, which are often supplied by a third party and in use by other

[17] Taken from: http://f1000research.com/articles/1-3/v1

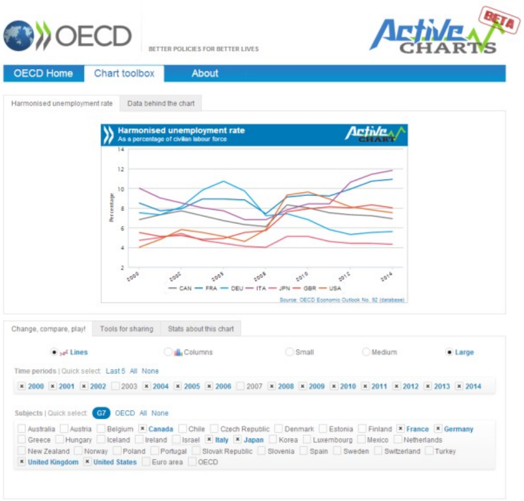
publishers, so making changes to these editorial systems would be difficult and time consuming.

There is also a question of how much dataset metadata reviewers expect to see on the journal site. Potentially, it would be less confusing for the reviewers and editorial staff to see the dataset metadata on the repository site, rather than mixed in with the article metadata. However, some metadata, e.g. a legend explaining the dataset, and details showing views, downloads, shares and citations are all information that are very useful for the article reader and also help to highlight how valuable the raw data itself may be.

Data Behind the Graph

One of the main aims of data publication is to make the research in articles more easily verifiable and reproducible by enabling the data underlying the article to become more visible. Increased accessibility to the data behind the graph also enables other researchers to make direct comparisons with previously published, or as yet unpublished, results. Where there is a plot in the journal article, clicking on it would redirect the reader to the subset of the data used to create that plot.

The OECD, as part of their ActiveCharts.org project, has provided an interactive site for re-mixing, visualising and sharing various statistics from the OECD's databases. Figure 11 shows an example of one of the graphs, while Figure 12 shows the raw data used to create it. The remixed graph can be shared and saved as a variety of formats for input into presentations, and can also be embedded into webpages.



● Lines ○ Columns ○ Small ○ Medium ● Large

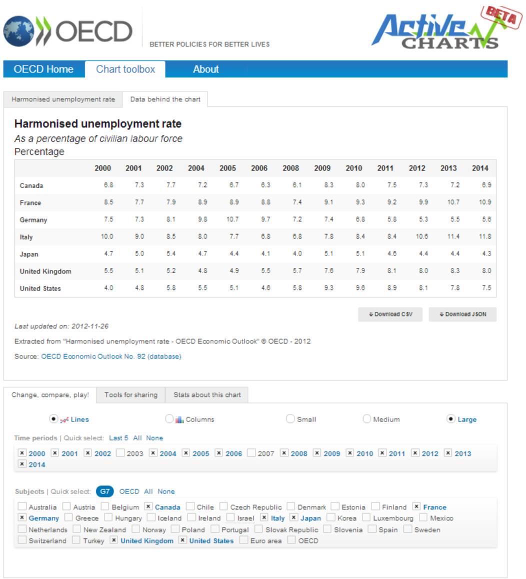
Time periods | Quick select: Last 5 All None

☒ 2000 ☒ 2001 ☒ 2002 □ 2003 ☒ 2004 ☒ 2005 ☒ 2006 □ 2007 ☒ 2008 ☒ 2009 ☒ 2010 ☒ 2011 ☒ 2012 ☒ 2013 ☒ 2014

Subjects | Quick select: G7 OECD All None

□ Australia □ Austria □ Belgium ☒ Canada □ Chile □ Czech Republic □ Denmark □ Estonia □ Finland ☒ France ☒ Germany
□ Greece □ Hungary □ Iceland □ Ireland □ Israel ☒ Italy ☒ Japan □ Korea □ Luxembourg □ Mexico □ Netherlands
□ New Zealand □ Norway □ Poland □ Portugal □ Slovak Republic □ Slovenia □ Spain □ Sweden □ Switzerland □ Turkey
☒ United Kingdom ☒ United States □ Euro area □ OECD

Figure 11. Active Chart created and displayed, featuring user selection tick boxes to display/hide data and re-plot it.[18]



Harmonised unemployment rate

As a percentage of civilian labour force

Percentage

	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

Last updated on: 2012-11-26

Extracted from "Harmonised unemployment rate - OECD Economic Outlook" © OECD - 2012

Source: OECD Economic Outlook No. 92 (database)

● Lines ○ Columns ○ Small ○ Medium ● Large

Time periods | Quick select: Last 5 All None

☒ 2000 ☒ 2001 ☒ 2002 □ 2003 ☒ 2004 ☒ 2005 ☒ 2006 □ 2007 ☒ 2008 ☒ 2009 ☒ 2010 ☒ 2011 ☒ 2012 ☒ 2013
☒ 2014

Subjects | Quick select: G7 OECD All None

□ Australia □ Austria □ Belgium ☒ Canada □ Chile □ Czech Republic □ Denmark □ Estonia □ Finland ☒ France
☒ Germany □ Greece □ Hungary □ Iceland □ Ireland □ Israel ☒ Italy ☒ Japan □ Korea □ Luxembourg □ Mexico
□ Netherlands □ New Zealand □ Norway □ Poland □ Portugal □ Slovak Republic □ Slovenia □ Spain □ Sweden
□ Switzerland □ Turkey ☒ United Kingdom ☒ United States □ Euro area □ OECD

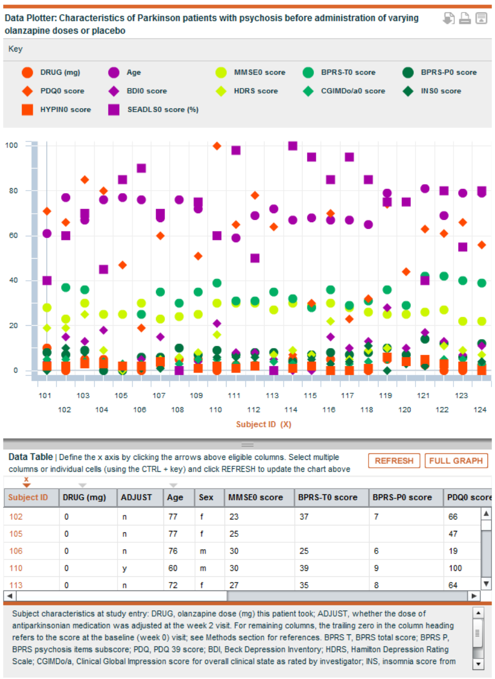
Figure 12. Data behind the graph shown in Figure 11.

The functionality in ActiveCharts.org could be integrated into journal paper webpages, although at this time embedding an active chart into a webpage only provides

[18] Retrieved from: http://activecharts.org/share/a7dd3bae149b2aba5b8f0d895e00d364

the published chart – the user has to click on the link in the shared chart to get back to ActiveCharts.org, where they can then modify the visualisation of the data.

Similar technology has been developed recently by F1000Research, who have released a beta version of a data plotting tool, enabling readers to re-plot raw spreadsheet data, changing the x and y-axis as appropriate. Figure 13 shows the author's plot and the raw data, and Figure 14 shows an example re-plot that a reader or referee could create by changing the x-axis and only plotting a couple of the metrics on the y-axis, thus enabling the data to be 'played with' on the fly within the article itself.



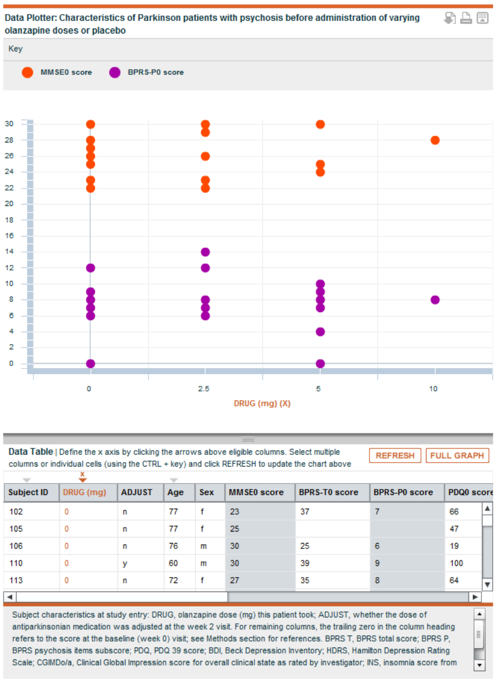
Data Plotter: Characteristics of Parkinson patients with psychosis before administration of varying olanzapine doses or placebo

Data Table | Define the x axis by clicking the arrows above eligible columns. Select multiple columns or individual cells (using the CTRL + key) and click REFRESH to update the chart above

Subject ID	DRUG (mg)	ADJUST	Age	Sex	MMSE0 score	BPRS-T0 score	BPRS-P0 score	PDQ0 score
102	0	n	77	f	23	37	7	66
105	0	n	77	f	25			47
106	0	n	76	m	30	25	6	19
110	0	y	60	m	30	39	9	100
113	0	n	72	f	27	35	8	64

Subject characteristics at study entry: DRUG, olanzapine dose (mg) this patient took; ADJUST, whether the dose of antiparkinsonian medication was adjusted at the week 2 visit. For remaining columns, the trailing zero in the column heading refers to the score at the baseline (week 0) visit; see Methods section for references. BPRS T, BPRS total score; BPRS P, BPRS psychosis items subscore; PDQ, PDQ 39 score; BDI, Beck Depression Inventory; HDRS, Hamilton Depression Rating Scale; CGIMDo/a, Clinical Global Impression score for overall clinical state as rated by investigator; INS, insomnia score from

Figure 13. F1000Research data plotting tool showing the raw data and the author graph of that data (Nichols et al., 2013).



Data Plotter: Characteristics of Parkinson patients with psychosis before administration of varying olanzapine doses or placebo

Data Table | Define the x axis by clicking the arrows above eligible columns. Select multiple columns or individual cells (using the CTRL + key) and click REFRESH to update the chart above

Subject ID	DRUG (mg)	ADJUST	Age	Sex	MMSE0 score	BPRS-T0 score	BPRS-P0 score	PDQ0 score
102	0	n	77	f	23	37	7	66
105	0	n	77	f	25			47
106	0	n	76	m	30	25	6	19
110	0	y	60	m	30	39	9	100
113	0	n	72	f	27	35	8	64

Subject characteristics at study entry: DRUG, olanzapine dose (mg) this patient took; ADJUST, whether the dose of antiparkinsonian medication was adjusted at the week 2 visit. For remaining columns, the trailing zero in the column heading refers to the score at the baseline (week 0) visit; see Methods section for references. BPRS T, BPRS total score; BPRS P, BPRS psychosis items subscore; PDQ, PDQ 39 score; BDI, Beck Depression Inventory; HDRS, Hamilton Depression Rating Scale; CGIMDo/a, Clinical Global Impression score for overall clinical state as rated by investigator; INS, insomnia score from

Figure 14. An example re-plot of the data in Figure 13 using the F1000-Research data plotting tool.

GDJ is a data journal, and so is primarily concerned with publishing information about datasets without the need for drawing conclusions from the data. For this reason, the figures in GDJ are likely to be examples of representative sections of the dataset being published. At a simple level, for a time-series graph representing, for example, a single day of measurements, it would be possible for the user to click through and download the file containing that day's worth of measurements. This is assuming that the repository can offer download of the data at that resolution, which may not always be the case. The link to that particular file would have to be managed carefully, as it may not be appropriate to assign a DOI to a single file. This also ties in with the issue of citation granularity, where the data behind the graph becomes a citeable entity in its own right, though related to the larger cited dataset. This issue is still being discussed by the community, and as yet, no clear guidance on how to deal with it has emerged.

For most other journals, figures are more likely to be generated from processed data stored on the researcher's local workstation, and will probably not be ingested into a data repository in a formal way, although some journals (e.g. F1000Research) are including the raw data behind all the results published in their articles. Some repositories (e.g. figshare) allow files to be deposited which could contain the data used to generate a single figure, and assign DOIs to those files. In that case, it would be possible to link from the figure to the data file. It is also possible to 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 much more closely bound to, the articles themselves in the form of supporting information.

This method of cross-linking relies on authors being willing and able to submit the exact data subsets they used to create each figure, and therefore may involve additional work, both in producing the subsets, but also archiving them properly (though this may vary according to the journal). There would need to be a general consensus between authors and publishers of the benefits of this additional work towards publication.

Figures submitted to a journal for publication will be transformed by the publication process as a matter of course.

Recommendations and Conclusions

This report has outlined a number of potential methods for cross-linking between journals and repositories for the purposes of data publication, and we have also investigated how they might be implemented in a journal. The first, linking using DOIs, is the most established and has been demonstrated in Geoscience Data Journal and others. Of the many cross-linking examples detailed here, their primary benefit to the reader is in connecting the reader directly to the data that underlies the research publication. This may be done in a variety of different ways, with different costs and benefits associated with each.

There are three main recommendations forthcoming from this work:

1. **Standardisation of metadata:** For cross-linking to be scalable across multiple journals and data repositories, automatic processes for the linking and sharing of metadata need to be developed. These processes require common standards that are applicable across a wide range of research domains. The project therefore recommends the use of 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.

2. **Use of DOIs and data citation:** As DOIs are persistent and actionable links, and are commonly used across the majority of publishers for linking between papers, it is recommended to use them for linking articles to data. This linking should be done in the context of a formal data citation, where other information about the dataset is given, including the creators, title, publishers and date of publication. This project recommends the DataCite citation structure given in the DataCite metadata schema v3.0 (DataCite, 2013), although 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, and the journal's author guidelines should be updated to request authors to cite the datasets used in their article (preferably using DOIs).

3. **Role of a centralised, third-party registry:** There is a role for a centralised, third-party registry and metadata broker in data publication to 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. We recommend that this be investigated through the Publishing Data Interest Group of the Research Data Alliance.

Acknowledgements

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

References

Callaghan, S.A., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., … Whyte, A. (2013). Processes and procedures for data publication: A case study in the geosciences. *International Journal of Digital Curation, 8*(1), 193-203. doi:10.2218/ijdc.v8i1.253

DataCite. (2013). *DataCite metadata schema for the publication and citation of research data: Version 3.0.* doi:10.5438/0008

Duerr, R.E., Downs, R.R., Tilmes, C., Barkstrom, B., Lenhardt, W.C., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: An assessment and recommendations. *Earth Science Informatics, 4*(3), 139-160. doi:10.1007/s12145-011-0083-6

Keall, B. (2012). *How Elsevier's Article of the Future supports researchers in the digital era.* Paper presented at the EGU General Assembly, Vienna, Austria. Abstract retrieved from http://adsabs.harvard.edu/abs/2012EGUGA..1412759K

Lawrence, R. (2012). New figshare widget to provide easy access to previews of data files on F1000 Research [Web log post]. Retrieved from F1000Research Blog: http://blog.f1000research.com/2012/07/10/new-figshare-widget-gives-previews-of-data-on-f1000-research/

Nichols, M.J., Hartlein, J.M., Eicken, M.G., Racette, B.A., & Black, K.J. (2013). A fixed-dose randomized controlled trial of olanzapine for psychosis in Parkinson disease [v1; ref status: indexed, http://f1000r.es/1au]. *F1000Research 2013, 2,* article 150. doi:10.12688/f1000research.2-150.v1