Data journals: building partnerships between publishers and data centres

Sarah Callaghan, Graham Parton, the PREPARDE project team and the NERC Data Citation and Publication Project Team
sarah.callaghan@stfc.ac.uk
@sorcha.ni

NCAS Research Forum: Delivering world leading research and innovation for industrial advantage
Tuesday 7 May 2013
Cedar Court Grand Hotel & Spa, York
Who are we and why do we care about data?

The UK’s Natural Environment Research Council (NERC) funds six data centres which between them have responsibility for the long-term management of NERC’s environmental data holdings.

We deal with a variety of environmental measurements, along with the results of model simulations in:

- Atmospheric science
- Earth sciences
- Earth observation
- Marine Science
- Polar Science
- Terrestrial & freshwater science, Hydrology and Bioinformatics
Even the Chancellor says data’s important!

“The next generation of scientific discovery will be data-driven discovery……”

“We need to make sure we capture value from this mass of data – both for economic growth and for social advances, such as better health.”

“This requires a transformation in data management”

Speech by the Chancellor of the Exchequer, Rt Hon George Osborne MP, to the Royal Society – 9 Nov 2012

Thanks to Jonathan Tedds (University of Leicester)
A key part of the scientific method is that it should be reproducible – other people doing the same experiments in the same way should get the same results.

Unfortunately observational data is not reproducible (unless you have a time machine!)

The way data is organised and archived is crucial to the reproducibility of science and our ability to test conclusions.

This is often the only part of the process that anyone other than the originating scientist sees.

We want to change this.
Journals have always published data…

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

…but datasets have gotten so big, it’s not useful to publish them in hard copy anymore

The Scientific Papers of William Parsons, Third Earl of Rosse 1800-1867
Reasons for citing and publishing data

• **Pressure** from (UK) government to make data from publicly funded research available for free.
  • Scientists want attribution and **credit** for their work
  • Public want to know what the scientists are doing
  • Good for the **economy** if new industries can be built on scientific data/research

• **Research funders** want reassurance that they’re getting value for money
  • Relies on peer-review of science publications (well established) and data (starting to be done!)

• Allows the wider research community and **industry** 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

What it all comes down to:

Encourage and provide credit to researchers and institutions for managing and disseminating their data properly.

Making data available is good for science and good for everyone - including UK PLC!
Why not just share the data?

Benefits of sharing:
• Ability to discover and reuse data which has already been collected
• Avoid redundant data collection
• Save time and money
• Provide opportunities for collaboration.

Research funders are keen to encourage data sharing.

For the most part, scientists are happy to share other scientists’ data, but...
Data may mean the difference between getting a grant and not.

There is (currently) no universally accepted mechanism for data creators to obtain academic credit for their dataset creation efforts.

Creators (understandably) prefer to hold the data until they have extracted all the possible publication value they can.

This behaviour comes at a cost for the wider scientific community.

Reframing “sharing” as “publication” might encourage scientists to be more open with their data.
Serving, citing and publishing data

0. Serving of data sets (Data centres)

1. Data set Citation (Everyone!)

2. Publication of data sets (Journal publishers)

This involves the peer-review of data sets, and gives “stamp of approval” associated with traditional journal publications. Can’t be done without effective linking/citing of the data sets.

Citation needs **permanent and unambiguous global identifiers**. Citing something means that you want to get the same thing back when you de-reference the citation - which is why we’re using DOIs.

This is what data centres do as our day job – take in data supplied by scientists and make it available to other interested parties. We have many ways to locate and identify the data in our archive. Note that the data can and does change!
How to publish data

• Stick it up on a webpage somewhere
  • Issues with stability, persistence, discoverability…
  • Maintenance of the website

• Put it in the cloud
  • Issues with stability, persistence, discoverability…

• Attach it to a journal paper and store it as supplementary materials
  • Journals not too keen on archiving lots of supplementary data, especially if it’s large volume.

• Put it in a disciplinary/institutional repository

• Write a data article about it and publish it in a data journal

By David Fletcher
“Publishing” versus “publishing” and “Open” versus “Closed”

Distinction between:

**Publishing** = publishing after some formal process which adds value for the consumer:
- e.g. PloS ONE type review, or
- EGU journal type public review, or
- More traditional peer review, and
- provides commitment to persistence

And **publishing/serving** = making available for consumption (e.g. on the web)
We want to:

Encourage scientists to move away from storing their data on CDs in their locked filing cabinets...

....or on hard disks with no backups....

And get them to put their data in a place where it’ll be archived and looked after for the future properly…

…where it can be shared/made available/published for the benefit of other researchers/general public/policy makers
Partnering with journal publishers to publish data

The scientific quality of a dataset has to be evaluated by peer-review by scientists with domain knowledge. This peer-review process has already been set up by academic publishers, so it makes sense to collaborate with them for peer-review publishing of data.

Can cite using URLs, but we’ve realised that people don’t trust URLs. We’re loading DOIs with more meaning than them simply being a persistent identifier – using them to signify completeness and technical quality of the dataset.

The day job – take in data and metadata supplied by scientists (often on an ongoing basis). Make sure that there is adequate metadata and that the data files are appropriate format. Make it available to other interested parties.

0. Serving of data sets (Data centres)

1. Data Set Citation (Everyone!)

2. Publication of data sets (Journal publishers)
Now we can cite our datasets using DOIs, we can give academic credit to those scientists who get cited – making them more likely to give us good quality data to archive.

Publication – and scientific peer-review – is the next step

We are working with the Royal Meteorological Society and Wiley-Blackwell to operate a new data journal, the Geoscience Data Journal

GDJ is an online-only, Open Access journal, publishing short data papers cross-linked to – and citing – datasets that have been deposited in approved data centres and awarded DOIs.

Other data journals already exist – see a list (in no particular order) at:
http://proj.badc.rl.ac.uk/prepared/blog/DataJournalsList
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).
Scientific Data is a new open-access, online-only publication for descriptions of scientifically valuable datasets. It introduces a new type of content called the Data Descriptor, which will combine traditional narrative content with curated, structured descriptions of research data, including detailed methods and technical analyses supporting data quality.
Publishing data for the scholarly record

- Scientific journal publication mainly focuses on the analysis, interpretation and conclusions drawn from a given dataset.

- Examining the raw data that forms the dataset is more difficult, as datasets are usually stored in digital media, in a variety of (proprietary or non-standard) formats.

- Peer-review is generally only applied to the methodology and final conclusions of a piece of work, and not the underlying data itself. But if the conclusions are to stand, the data must be of good quality.

- A process of data publication, involving peer-review of datasets would be of benefit to many sectors of the academic community.

Most scientists regarded the new streamlined peer-review process as ‘quite an improvement.’

http://libguides.luc.edu/content.php?pid=5464&sid=164619
**The traditional online journal model**

1) Author prepares the paper using word processing software.

2) Author submits the paper as a PDF/Word file.

3) Reviewer reviews the PDF file against the journal’s acceptance criteria.

**Overlay journal model for publishing data**

1) Author prepares the data paper using word processing software and the dataset using appropriate tools.

2a) Author submits the data paper to the journal.

2b) Author submits the dataset to a repository.

3) Reviewer reviews the data paper and the dataset it points to against the journals acceptance criteria.
What is a data article?

A data article describes a dataset, giving details of its collection, processing, software, file formats, etc., without the requirement of novel analyses or groundbreaking conclusions.

- the *when, how and why* data was collected and what the data-product is.
PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences

Funded by JISC

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
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
Live Data Paper in Geoscience Data Journal!

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
**GSB 20.7GHz slant path radio propagation measurements, Chilbolton site**

**General Info**
- **Title:** GSB 20.7GHz slant path radio propagation measurements, Chilbolton site
- **Type:** Activity
- **Sub-Type:** Deployment
- **Publications Citation:**
  - **URL:** http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM.dep_11902119479621181

**Summary**
The GSB (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, with signal attenuation sampled once per second.

**This dataset is cited in:**

**Author**
- **Name:** Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research. [S. A. Callaghan, J. Weight, C. J. Walden, J. Agnew and S. Ventouras]

**Online References**
- **Relation:** Apply for access
  - **Title:** Apply for data from Chilbolton
- **Relation:** Download
  - **Title:** Directory for GSB data from Chilbolton
  - **DOI:** 10.1002/gjd3.2
- **Relation:** Documentation
  - **Title:** Data article in Geoscience Data Journal. DOI: 10.1002/gjd3.2

**Associated Data**
- **Type:** Chilbolton, GBS receiver
- **Activity:** Chilbolton Facility for Atmospheric and Radio Research (CPARR)
- **Observation Station:** Chilbolton Facility for Atmospheric and Radio Research (CPARR), UK

**Reference to Data Article**
- Clickable link to Data Article
Working with Elsevier for publication to data linking

Data journals are a special case of journal publisher/data centre interactions.

There is still the need to link to data (held in repositories) from journal papers that mention/cite that data.

We’re working with Elsevier to do just that.

Elsevier have updated their Guide for Authors text.
How data and articles are linked

There are several ways in which we support interlinking of articles and data:

- **Referencing data in your article through tagging identifiers or accession numbers:** If your article contains relevant unique identifiers or accession numbers linking to information on genes, proteins, diseases, etc. or structures deposited in public databases, and you would like your article to link to that data, please identify these entities in the following way:

  database abbreviation: data identifier

  For example, "PDB: 1TUP" to identify the protein with accession number "1TUP" in the Protein Data Bank (PDB). Please bear in mind that an error in a letter or number will result in a dead link in the article. Database abbreviations and further examples can be found in the listing of supported databases.

- **Data DOI's:** Elsevier supports Data DOI's as persistent identifiers for scientific data. If you include a data DOI in your article, it will automatically turn into a link to your data on ScienceDirect.

- **Linked data repository banners on ScienceDirect:** Elsevier collaborates with selected data repositories to show banner links next to relevant articles on ScienceDirect. This linking system requires that the data repository maintains accurate records of associations between articles and data sets. What you need to do as an author to support this type of linking depends on the data repository; see links to more information in the supported databases section.

- **Data visualization and integration applications:** In close collaboration with selected data repositories, Elsevier has developed a number of data-integration and visualization applications that are shown next to the article on ScienceDirect, e.g. the Protein Viewer (with PDB), the PANGAEA data visualization tool, and the Genome Viewer (with NCBI). These applications build further on tagged entities or banner links to visualize data and integrate it into the online reading experience.

From: [http://www.elsevier.com/about/content-innovation/database-linking#about-database-linking](http://www.elsevier.com/about/content-innovation/database-linking#about-database-linking)
### Earth, Environmental & Oceanographic Data

<table>
<thead>
<tr>
<th>Data Repository</th>
<th>How articles and data are linked</th>
<th>More information</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGS GeoScenic</td>
<td>Authors should specify BGS GeoScenic numbers, e.g. GeoScenic: P603281.</td>
<td>• BGS GeoScenic homepage</td>
</tr>
<tr>
<td>EarthChem</td>
<td>EarthChem banners will be shown on ScienceDirect when the repository has data for the article. Data is extracted from the literature by curators.</td>
<td>• EarthChem homepage</td>
</tr>
<tr>
<td>Marine Geoscience Data System (MGDS)</td>
<td>MGDS banners will be shown on ScienceDirect when the repository has data for the article.</td>
<td>• MGDS homepage</td>
</tr>
<tr>
<td>Natural Environment Research Council (NERC), including BADC, BODC, EIDC, and NGDC.</td>
<td>Authors should include data DOI’s in their manuscript.</td>
<td>• NERC Data Centres</td>
</tr>
<tr>
<td>PANGAEA</td>
<td>Data integration application on ScienceDirect opens automatically for relevant articles.</td>
<td>• PANGAEA homepage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• PANGAEA application</td>
</tr>
<tr>
<td>System for Earth Sample Registration (SESAR), registry for International Geo Sample Numbers (IGSN)</td>
<td>Authors should specify IGSN numbers, e.g. IGSN: HRV003M16.</td>
<td>• SESAR homepage</td>
</tr>
<tr>
<td>Woods Hole Open Access Server (WHOAS)</td>
<td>WHOAS banners will be shown on ScienceDirect when the repository has data for the article.</td>
<td>• WHOAS homepage</td>
</tr>
</tbody>
</table>

http://www.elsevier.com/about/content-innovation/database-linking#supported-data-repositories

NERC data centres are listed in Elsevier’s list of supported data repositories.
Elsevier working with NGDC to link through accession numbers

Hyperlinked GeoScenic Accession Numbers in the article main text (e.g. “GeoScenic: P100659”) – tagged by authors

Available for all Elsevier geology journals

Thanks to Bethan Keall (Elsevier)
Linking with data sets

The journal would like to encourage authors to link to relevant data sets underpinning their research publication which are archived in recognised data centres, such as those of the Natural Environment Research Council (NERC). The preferred way to do this is by adding the DOI of the data set into the manuscript. Elsevier will turn these DOI’s into links in the online article, making it easy for readers to find data pertinent to the published article. Elsevier would also like to encourage authors to deposit the data that supports their publication in an appropriate data archive.

http://strangefunny.com/research-cat-says/
Observations of Fukushima fallout in Great Britain

N.A. Beresford1, C.L. Barnett2, B.J. Howard3, D.C. Howard4, C. Wells5, A.N. Tyler6, S. Bradley7, D. Copplestone8

1 Centre for Ecology & Hydrology, Lancaster Environment Centre, Library Av., Bingley, Lancaster LA1 4AP, United Kingdom
2 Institute of Biological and Environmental Sciences, School of Natural Sciences, University of Stirling, Stirling FK9 4LA, United Kingdom

http://dx.doi.org/10.1016/j.jenvrad.2011.12.008, How to Cite or Link Using DOI

Abstract

Following the Fukushima accident in March 2011, grass samples were collected from 42 sites around Great Britain during April 2011. Iodine-131 was measurable in grass samples across the country with activity concentrations ranging from 10 to 65 Bq kg\(^{-1}\) dry matter. Concentrations were similar to those reported in other European countries. Rainwater and some foodstuffs were also analysed from a limited number of sites. Of these, \(^{131}\)I was only detectable in sheep’s milk (c. 2 Bq kg\(^{-1}\)). Caesium-134, which can be attributed to releases from the Fukushima reactors, was detectable in six of the grass samples (4–8 Bq kg\(^{-1}\) dry matter); \(^{129}\)Cs was detected in a larger number of grass samples although previous release sources (atmospheric weapons tests and the 1986 Chernobyl and 1987 Windscale accidents) are likely to have contributed to this.

Highlights

- Grass samples from across Great Britain were sampled and analysed following releases from the Fukushima accident.
- Iodine-131 was detectable, at low levels, in grass samples from throughout the UK.

Table 1

<table>
<thead>
<tr>
<th>Sample Type</th>
<th>Activity Concentration (Bq kg(^{-1}) dry matter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grass</td>
<td>10 - 65</td>
</tr>
<tr>
<td>Sheep’s milk</td>
<td>2</td>
</tr>
<tr>
<td>Rainwater</td>
<td>4 - 8</td>
</tr>
<tr>
<td>Foodstuffs</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>

References

...to the underlying dataset, using DOIs
The Data Citation Index on the Web of Knowledge platform provides a single point of access to quality research data from repositories across disciplines and around the world.

Through linked content and summary information, this data is displayed within the broader context of the scholarly research, enabling users to gain perspective that is lost when data sets or repositories are viewed in isolation. These connections allow researchers to efficiently access to an array of data across subjects and regions, providing a comprehensive picture of research output, to maximize research efforts and accurately assess importance.

For more than 50 years, Thomson Reuters has provided intelligent information to accelerate research, discovery, and innovation.
<table>
<thead>
<tr>
<th>REPOSITORY</th>
<th>DISCIPLINE</th>
<th>RESPONSIBLE ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaeological Data Service</td>
<td>Social Sciences</td>
<td>University of York</td>
</tr>
<tr>
<td>Array Express</td>
<td>Life Sciences</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>Association of Religion Data Archives</td>
<td>Arts and Humanities</td>
<td>Pennsylvania State University</td>
</tr>
<tr>
<td>Australian Antarctic Data Centre</td>
<td>Climatology</td>
<td>Australian Government; Department of Sustainability, Environment, Water, Population and Communities</td>
</tr>
<tr>
<td>Australian Data Archive</td>
<td>Social Sciences</td>
<td>Australian National University</td>
</tr>
<tr>
<td>British Antarctic Survey</td>
<td>Physical Sciences</td>
<td>Natural Environment Research Council</td>
</tr>
<tr>
<td>British Atmospheric Data Centre</td>
<td>Atmospheric Science</td>
<td>Natural Environment Research Council</td>
</tr>
<tr>
<td>British Geological Survey</td>
<td>Physical Sciences</td>
<td>Natural Environment Research Council</td>
</tr>
<tr>
<td>British Oceanographic Data Centre</td>
<td>Physical Sciences</td>
<td>Natural Environment Research Council</td>
</tr>
<tr>
<td>CoReudy</td>
<td>Life Sciences</td>
<td>National School of Agriculture</td>
</tr>
<tr>
<td>caNanocl</td>
<td>Life Sciences</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>ConGEM</td>
<td>Life Sciences</td>
<td>University of Helsinki</td>
</tr>
<tr>
<td>CEBS</td>
<td>Life Sciences</td>
<td>The National Institute of Environmental Health Sciences</td>
</tr>
<tr>
<td>Cell Centered Database</td>
<td>Neuroscience</td>
<td>University of California</td>
</tr>
<tr>
<td>Centre for Ecology and Hydrology</td>
<td>Physical Sciences</td>
<td>Natural Environment Research Council</td>
</tr>
<tr>
<td>Codex Sinaicicus</td>
<td>Arts and Humanities</td>
<td>The British Library/Leipzig University Library/ St. Catherine's Monastery/The National Library of Russia</td>
</tr>
<tr>
<td>CPLA</td>
<td>Life Sciences</td>
<td>University of Science and Technology of China</td>
</tr>
<tr>
<td>Crystallography Open Database</td>
<td>Physical Sciences</td>
<td>Vميلز University*</td>
</tr>
<tr>
<td>Disprot</td>
<td>Life Sciences</td>
<td>Indiana University School of Medicine/ Temple University</td>
</tr>
<tr>
<td>DrugBank</td>
<td>Life Sciences</td>
<td>University of Alberta*</td>
</tr>
<tr>
<td>Dryad</td>
<td>Life Sciences</td>
<td>National Evolutionary Synthesis Center</td>
</tr>
<tr>
<td>EcoGene</td>
<td>Life Sciences</td>
<td>University of Miami</td>
</tr>
<tr>
<td>eCrystals</td>
<td>Crystallography</td>
<td>University of Southampton</td>
</tr>
<tr>
<td>Emage</td>
<td>Life Sciences</td>
<td>Medical Research Council</td>
</tr>
<tr>
<td>EMDDB</td>
<td>Life Sciences</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>Esther</td>
<td>Life Sciences</td>
<td>French National Institute for Agricultural Research*</td>
</tr>
<tr>
<td>Eurostat</td>
<td>Social Sciences</td>
<td>European Union</td>
</tr>
<tr>
<td>Finnish Social Science Data Archive</td>
<td>Social Sciences</td>
<td>University of Tampera</td>
</tr>
<tr>
<td>Gene Expression Omnibus</td>
<td>Genetics</td>
<td>National Center for Biotechnology Information</td>
</tr>
</tbody>
</table>
|        |        |                                        | http://wokinfo.com/products_tools/multidisciplinary/dci/repositories/
What we’ve done and how we’ve done it

Data paper has been published in a data journal, linked via DOI to underlying dataset. Formal citations of datasets (also using DOIs) done in standard academic articles.

Can cite using URLs, but we’ve realised that people don’t trust URLs. We’re loading DOIs with more meaning than them simply being a persistent identifier – using them to signify completeness and technical quality of the dataset. We’re also looking at citation counts as metric for dataset impact.

The day job – take in data and metadata supplied by scientists (often on a ongoing basis). Make sure that there is adequate metadata and that the data files are appropriate format. Make it available to other interested parties.
• The NERC data centres now have the ability to mint DOIs and assign them to datasets in their archives. We have also produced:
  • guidelines for the data centre on what is an appropriate dataset to cite
  • guidelines for data providers about data citation and the sort of datasets we will cite
  • text in the NERC grants handbook telling grant applicants about data citation

• We’ve already had users coming to us requesting DOIs for their datasets.

• We’re progressing well with data publication through our partnership with Wiley-Blackwell, and discussions with Elsevier and Thompson-Reuters. NERC held datasets have been published in data journals and cited in papers.

• Still plenty of work to do! Not just mechanical processes (e.g. workflows, guidelines) but also changing the culture so that citing and publishing data is the norm.
Science not alchemy!

sarah.callaghan@stfc.ac.uk
@sorcha_ni
http://citingbytes.blogspot.co.uk/
data-publication@jiscmail.ac.uk

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

Guidelines on peer review for data:
Feedback to:
https://www.jiscmail.ac.uk/DATA-PUBLICATION

Page from alchemic treatise of Ramon Llull
(Beginning of the 16th century)