Environmental Data Archival: Practices and Benefits

Graham Parton
graham.parton@stfc.ac.uk

With many thanks to Dr Sarah Callaghan

Transmission, presentation and archiving of meteorological data
Quick into about CEDA

The Centre for Environmental Data Archival is responsible for the running of the following data centres:

- **British Atmospheric Data Centre**
  - NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.

- **The British Atmospheric Data Centre**

- **NERC Earth Observation Data Centre**
  - The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.

- **The UK Solar System Data Centre**
  - The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

- **IPCC Data Distribution Centre**
  - The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.
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
But…

Why archive data anyway?
The “Scientific” Method

1. Define/Identify the Problem
2. Form a Hypothesis
3. Make Observations
4. Test Hypothesis Perform Experiments
5. Organize and Analyze Data
6. Do Experiments and Observations Support Hypothesis?
7. Draw Conclusions
8. Communicate Results

Results critiqued by peers?

Use method/data in subsequent research?

Need to refine hypothesis and test new hypothesis?

Yes

No

Scientific method built on:

- **reproducibility** and
- **transparency** of method and results

Allows peers to critique
- method – reasoned approach to collect data
- results - analysed and synthesised= plots
- conclusions – results correctly interpreted

Published in peer reviewed literature…

Thanks to Nik Papageorgiou: [http://upmic.wordpress.com/](http://upmic.wordpress.com/)
Traditionally: Everything in Journals. Including data

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

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Object</th>
<th>Magnitude</th>
<th>Ang. Pos.</th>
<th>No. of minutes</th>
<th>Distance</th>
<th>No. of minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>30th April</td>
<td>N a</td>
<td>12 02</td>
<td>3</td>
<td>607</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>30th April</td>
<td>N a b</td>
<td>13 42</td>
<td>3</td>
<td>607</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>30th April</td>
<td>N a c</td>
<td>13 39</td>
<td>3</td>
<td>607</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>30th April</td>
<td>N a d</td>
<td>13 40</td>
<td>3</td>
<td>607</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

The Scientific Papers of William Parsons, Third Earl of Rosse 1800-1867
New paradigm: NOT Everything in Journals

...but datasets are now very large volume:

CERN ~15 petabytes data produced annually

DNA ~ 1 exabyte of genome data by 2014

Climate change -

CMIP5: Fifth Coupled Model Intercomparison Project

Produced data underpinning next IPCC report
Simulations:
~90,000 years
~60 experiments
~20 modelling centres (from around the world) using
~30 major(*) model configurations
~2 million output “atomic” datasets
~10's of petabytes of output
~2 petabytes of CMIP5 requested output
~1 petabyte of CMIP5 “replicated” output
Which are replicated at a number of sites (including ours)

Of the replicants:
~ 220 TB decadal
~ 540 TB long term
~ 220 TB atmosphere-only

~80 TB of 3hourly data
~215 TB of ocean 3d monthly data
~250 TB for the cloud feedbacks
~10 TB of land-biochemistry (from the long term experiments alone)
So, SCALE of data means it's no longer possible to publish data anymore...

So what is under threat?
Reproducibility and transparency

How do we answer this? .... Do we even care?

What are our options?

Are synthesis/plots sufficient?
Just carry on with small scale datasets alone?
Just reproduce the data?....
Or just get on and archive the stuff?
Creating a dataset is hard work!

"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com
Why should I bother putting my data into a repository?

"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com
It’s ok, I’ll just do regular backups

These documents have been preserved for thousands of years! But they’ve both been translated many times, with different meanings each time.

Data Preservation is not enough, we need Active Curation to preserve Information
Benefits of archiving data

1. It costs time, effort and money to create it
2. Not all data are reproducible!
3. There is added benefit too in data re-use

CORRAL Project:
UK Colonial Registers and Royal Navy Logbooks
Initial costs of collecting the data in C19th
300 logbooks producing some
40,000 images

Useful for historical researchers & climate scientists
Bonuss material: Data Services
The research data lifecycle

Researchers are used to creating, processing and analysing data.

Data repositories generally have the job of preserving and giving access to data.

Third parties, or even the original researchers will reuse the data.

See http://data-archive.ac.uk/create-manage/life-cycle for more detail.
What is a Dataset?

DataCite’s definition (http://www.datacite.org/sites/default/files/Business_Models_Principles_v1.0.pdf):

Dataset: "Recorded information, regardless of the form or medium on which it may be recorded including writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow, charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data." (from the U.S. National Institutes of Health (NIH) Grants Policy Statement via DataCite's Best Practice Guide for Data Citation).

A dataset is something that is:
• The result of a defined process
• Scientifically meaningful
• Well-defined (i.e. clear definition of what is in the dataset and what isn’t)
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

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.

But if we publish the data, precedence is established and credit is given!
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)
What do we need to support publishing data?

0. Serving of data sets (Data centres)

1. Data Set Citation (Everyone!)

2. Publication of data sets (Journal publishers)

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.
Citing Data

- We already have a working method for linking between publications which is:
  - commonly used
  - understood by the research community
  - used to create metrics to show how much of an impact something has (citation counts)
  - applied to digital objects (digital versions of journal articles)

- We can extend citation to other things like:
  - data
  - code
  - multimedia

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

We using digital object identifiers (DOIs) as part of our dataset citation because:

- They are actionable, interoperable, persistent links for (digital) objects
- Scientists are already used to citing papers using DOIs (and they trust them)
- Academic journal publishers (e.g. Nature) are starting to require datasets be cited in a stable way, i.e. using DOIs.
- The British Library and DataCite approached us to pilot citing data using DOIs – and we’ve developed a good working relationship
What sort of data can we/will we assign a DOI to?

Dataset has to be:
• Stable (i.e. not going to be modified)
• Complete (i.e. not going to be updated)
• Permanent – by assigning a DOI we’re committing to make the dataset available for posterity
• Good quality – by assigning a DOI we’re giving it our data centre stamp of approval, saying that it’s complete and all the metadata is available

When a dataset is cited that means:
• There will be bitwise fixity
• With no additions or deletions of files
• No changes to the directory structure in the dataset “bundle”

A DOI should point to a html representation of some record which describes a data object – i.e. a landing page.

Upgrades to versions of data formats will result in new editions of datasets.
GBS 20.7GHz slant path radio propagation measurements, Chilbolton site

Summary
The GBS (Global Broadcast Service) dataset is a series of radio attenuation measurements made at three sites in the UK: Chilbolton and Spathroth, 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 the following:

This dataset is cited in:

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

Online References
Relation: Title
Apply for access: Data Directory for GBS Data from Chilbolton
Download: Data Directory for GBS Data from Chilbolton
Documentation: Data article in Geoscience Data Journal doi:10.1002/gj352

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
How ISIS cite their data

Dataset citation

Investigation title: Responsive polymer brushes grafted from gold nanoparticles: interaction with surfactant.

Creator: Titmarsh, S
Creator: Hu, H

DOI: 10.5286/ISIS.E.24079886

Data of Experiment: Sat May 01 11:45:33 BST 2010

Publisher: STFC ISIS Facility

Data format: RAW/Nexus

Data Citation

The recommended format for citing this dataset in a research publication is as: [author], [date], [title], [publisher], [doi]

For example:


Abstract

We will use SANS to characterize dispersions of gold nanoparticles functionalized with poly(dimethyl aminoethyl methacrylate) brushes grafted from 25 nm gold nanoparticle interfaces by surface-initiated ATRP. We will characterize the response of the polymer layers to pH and at low pH, where the polymer is ionically charged, to electrolyte concentration. Our main focus however will be to determine the nature of the surfactant aggregates formed in the brushes at concentrations of surfactant that lead to peaks in the surface tension of the dispersions at pH=3 and 10 and also lead to the formation of multilayer structures in equivalent brushes formed at planar interfaces. We will measure under particle-matching contrast, for dispersions bearing hydrogenous pDMAEMA and particle-matching DMAEMA with particle-matching SDS and hydrogenous SDS respectively.
A project report on the end of a DataCite DOI?
What else can be on the end of a DataCite DOI? Educational exhibits.
What else can be on the end of a DataCite DOI? Images
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.

http://libguides.luc.edu/content.php?pid=5464&sid=164619
More of that later… for now…

The archival process…
The Digital Curation Centre’s Curation Lifecycle Model provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle.

http://www.dcc.ac.uk/resources/curation-lifecycle-model
"an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community."

Where

"The information being maintained is deemed to need Long Term Preservation, even if the OAIS itself is not permanent. Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community".
Open Archival Information System (OAIS)

SIP = Supplier Information Product
AiP = Archive Information Product
DIP = Dataset Information Product
OAIS: Responsibilities

- Determine and continue to liaise with Designated User Community.
- Negotiate and accept data from Data Providers.
- Ensure that data, etc. are independently understandable.
- Make preserved data, etc. available.
- Ensure Data, etc. are preserved.
OAIS: Functions

- Data Management
- Ingest
- Archival Storage
- Administration
- Access
- Preservation Planning
- Common Services
Curation Problems

stories from the coal face…
Sometimes other people don’t get it.

"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com
Archiving can sometimes produce mixed feelings

"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com
Timeliness!

Thanks to Nik Papageorgiou: http://upmic.wordpress.com/
And some ways to overcome those problems...
Data Preparation

- Data Management Plans - agreed with PI/Programme SC
  - Delivery schedules
  - Conditions of Use/Licensing
  - Responsibilities of data providers and CEDA
  - Project specific requirements
  - 3rd party data requirements
  - shared group workspace
- Support suppliers in data preparation
  - Formats, metadata conventions
- File naming and archive structure
- Capture supporting documentation
  - (formats, calibration information, flight logs, etc.)
- Set up ingest routes
Good data and metadata formats

- Ensures future users can open data files
  - How future proof is an Excel spread sheet?
- Permits metadata harvesting from the data
- Generic extraction/processing tools for the data
- Can guarantee un-ambiguous content
### Data Preparation - File structure

Take the Bad Data Challenge…. File “sw010203”

<table>
<thead>
<tr>
<th>Value1</th>
<th>Value2</th>
<th>Value3</th>
<th>Value4</th>
<th>Value5</th>
<th>Value6</th>
<th>Value7</th>
<th>Value8</th>
<th>Value9</th>
<th>Value10</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.31</td>
<td>155.3</td>
<td>3.92</td>
<td>136.1</td>
<td>5.15</td>
<td>140.2</td>
<td>4.23</td>
<td>137.1</td>
<td>4.75</td>
<td>150.2</td>
</tr>
<tr>
<td>4.35</td>
<td>146.5</td>
<td>4.52</td>
<td>138.0</td>
<td>4.83</td>
<td>153.7</td>
<td>5.40</td>
<td>145.8</td>
<td>4.63</td>
<td>141.0</td>
</tr>
<tr>
<td>4.31</td>
<td>143.3</td>
<td>4.58</td>
<td>157.0</td>
<td>4.94</td>
<td>141.7</td>
<td>4.65</td>
<td>143.1</td>
<td>4.63</td>
<td>143.0</td>
</tr>
<tr>
<td>5.42</td>
<td>148.5</td>
<td>4.92</td>
<td>140.4</td>
<td>4.04</td>
<td>146.7</td>
<td>3.92</td>
<td>151.5</td>
<td>5.02</td>
<td>135.3</td>
</tr>
<tr>
<td>4.65</td>
<td>152.3</td>
<td>4.31</td>
<td>168.8</td>
<td>3.79</td>
<td>145.3</td>
<td>5.92</td>
<td>152.9</td>
<td>5.02</td>
<td>145.8</td>
</tr>
<tr>
<td>4.79</td>
<td>144.1</td>
<td>4.60</td>
<td>147.5</td>
<td>5.33</td>
<td>150.1</td>
<td>4.81</td>
<td>141.0</td>
<td>6.02</td>
<td>146.9</td>
</tr>
<tr>
<td>4.42</td>
<td>142.5</td>
<td>4.58</td>
<td>133.4</td>
<td>4.35</td>
<td>150.5</td>
<td>4.96</td>
<td>149.8</td>
<td>5.56</td>
<td>143.4</td>
</tr>
<tr>
<td>5.19</td>
<td>141.6</td>
<td>4.40</td>
<td>142.4</td>
<td>4.10</td>
<td>152.6</td>
<td>5.02</td>
<td>134.0</td>
<td>4.94</td>
<td>142.9</td>
</tr>
<tr>
<td>5.38</td>
<td>141.5</td>
<td>5.88</td>
<td>144.8</td>
<td>6.00</td>
<td>140.1</td>
<td>4.75</td>
<td>158.3</td>
<td>5.08</td>
<td>148.1</td>
</tr>
<tr>
<td>4.27</td>
<td>150.8</td>
<td>4.69</td>
<td>138.8</td>
<td>5.71</td>
<td>144.0</td>
<td>5.21</td>
<td>138.8</td>
<td>5.00</td>
<td>132.4</td>
</tr>
</tbody>
</table>

What are these data? Guess surface winds, but on what day?  
What are the units? Any convention?  
How do we read the file?  
Is this spatial or temporal data?... 1440 pairs of data in a file
netcdf simple {
  dimensions:
    latitude = 3 ;
    longitude = 2 ;
    time = UNLIMITED ; // (5 currently);
  variables:
    double time(time) ;
      time:standard_name = "time" ;
      time:units = "minutes since 1994-01-01 00:00:00" ;
      time:long_name = "time" ;
    float latitude(latitude) ;
      latitude:standard_name = "latitude" ;
      latitude:units = "degrees north" ;
      latitude:point_spacing = "even" ;
      latitude:long_name = "latitude" ;
    float longitude(longitude) ;
      longitude:standard_name = "longitude" ;
      longitude:units = "degrees east" ;
      longitude:point_spacing = "even" ;
      longitude:long_name = "longitude" ;
    float temp(time, latitude, longitude) ;
      temp:standard_name = "surface_temperature" ;
      temp:long_name = "Surface temperature in degrees C" ;
      temp:units = "deg_C" ;
      temp:_FillValue = 2.e+020f ;
      temp:valid_min = -80.0f ;
      temp:valid_max = 60.0f ;
      temp:comment = "This parameter may be erroneous." ;
}
Time for metadata?
Domain specific repositories can:

- Pick and choose what data to keep
- Ask for (and get) more detailed metadata
- Provide specific tools and services (visualisations, server-side processing, …)
- Deal with Big Data!

Libraries will need to:

- Pick up and manage/archive the long-tail data where there isn’t a domain repository
- Have generalised, widely applicable systems that can cope with subjects from astronomy to zoology
- Be prepared to cope with anything!
Summary and maybe conclusions?

• Data is important, and becoming more so for a far wider range of the population
• Conclusions and knowledge are only as good as the data they're based on
• Science is supposed to be reproducible and verifiable
• It's up to us as scientists to care for the data we've got and ensure that the story of what we did to the data is transparent
  • So we can use the data again
  • And so people will trust our results

• It's not an easy job – but someone's got to do it!
Thanks!

Any questions?
graham.parton@stfc.ac.uk
http://www.scoop.it/t/windgatherer/
sarah.callaghan@stfc.ac.uk
@sorcha_ni
http://citingbytes.blogspot.co.uk/

Metadata
It is generally agreed that we need methods to:

- define and document datasets of importance.
- augment and/or annotate data
- amalgamate, reprocess and reuse data

To do this, we need metadata – data about data

For example:
Longitude and latitude are metadata about the planet.
- They are artificial
- They allow us to communicate about places on a sphere
- They were principally designed by those who needed to navigate the oceans, which are lacking in visible features!

Metadata can often act as a surrogate for the real thing, in this case the planet.
Metadata for Discovery, Documentation, Definition

Dataset METADATA

DEFINITIONS

A: Archive. Usage metadata generated from (of about the data). Normally generated from internal metadata.

B: Browse. Context, Generic, semantic, including a summary of A-type, links to or embeds discipline specific (E-type).

C: Character and Citation. Post-fact annotations and citations both internal and external (trackback, comments).


E: Extra. Discipline specific metadata, may or may not be understood at all sites, e.g. SensorML, NumSim.

RELATIONSHIPS

Q: Defined & supported textural, semantic, and spatio-temporal queries.

S: Security

D: Discovery

C: Character and Citation

B: Browse

A: Archive

Observation Collection – Logical grouping of results, e.g. all data for a project

Observation – specific set of data – the **What** (date, time, description of result – the **Where, When, Who**)

Process – the **How**: Acquisition | Computation | Composite

Project – the **Why**

---

MOLES3
Data Journals
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.
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.
Data journals and scientific publication of data

- The NERC data centres (and other repositories) can now cite datasets using DOIs
  - we can give academic credit to those scientists who get cited
- 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/preparde/blog/DataJournalsList
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.
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
Dataset catalogue page (and DOI landing page) – again!

Reference to Data Article

Clickable link to Data Article
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)
<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: FG0328T.</td>
<td><a href="#">BGS GeoScenic homepage</a></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><a href="#">EarthChem homepage</a></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><a href="#">MGDS homepage</a></td>
</tr>
<tr>
<td>NERC Data Centres, including BADC, BODC, EDC, and NGDC.</td>
<td>Authors should include data DOI's in their manuscript.</td>
<td><a href="#">NERC Data Centres</a></td>
</tr>
<tr>
<td>PANGAEA</td>
<td>Data integration application on ScienceDirect opens automatically for relevant articles.</td>
<td><a href="#">PANGAEA homepage</a></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><a href="#">SESAR homepage</a></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><a href="#">WHOAS homepage</a></td>
</tr>
</tbody>
</table>

NERC data centres are listed in Elsevier’s list of supported data repositories.

[http://www.elsevier.com/about/content-innovation/database-linking#supported-data-repositories](http://www.elsevier.com/about/content-innovation/database-linking#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. Beresford*, C.L. Barnett*, B.J. Howard*, D.C. Howard*, C. Wells*, A.N. Tyler¹, S. Bradley¹, D. Copplestone²

* Centre for Ecology & Hydrology, Lancaster Environment Centre, Birdling Avenue, Lancaster LA1 4AP, United Kingdom
¹ Institute of Biological & Environmental Sciences, University of Stirling, Stirling FK9 4LA, United Kingdom

Observations of Fukushima fallout in Great Britain
Journal of Environmental Radioactivity, Volume 114, 2012, Pages 45-53
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 16 to 55 Bq kg⁻¹ dry matter. Concentrations were similar to those reported in other European countries. Rainwater and some footstuffs were also analysed from a limited number of sites. Of these, ¹³¹I was only detectable in sheep's milk (c. 2 Bq kg⁻¹). Cesium-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). Cesium-137 was detected in a larger number of grass samples although previous release sources (atmospheric weapons test and the 1966 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 country.

Related Science and Research Jobs

THE WELCOME TRUST/INDIA ALLIANCE
Early Career Fellowships in Biomedical Research in India

View full text
Purchase

Find it

View Record in Scopus

Citing and related articles

BD Falcon™ Multi-Flasks
Multiply your cell growth & multiply your productivity with BD Falcon™ Multi-Flasks!

The Sentinels
ESA is developing a family of five space missions dedicated to...
…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.
<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 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>BioMedResBank</td>
<td>Life Sciences</td>
<td>University of Wisconsin</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>CAMPAL</td>
<td>Life Sciences</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>cANARC</td>
<td>Life Sciences</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>CanGEM</td>
<td>Life Sciences</td>
<td>University of Helsinki</td>
</tr>
<tr>
<td>CeED</td>
<td>Life Sciences</td>
<td>The National Institute of Environmental Health Sciences</td>
</tr>
<tr>
<td>Cell Centred Database</td>
<td>Neurosciences</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 Sinaticus</td>
<td>Arts and Humanities</td>
<td>The British Library, Lehigh 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>Vienna University*</td>
</tr>
<tr>
<td>DrugBank</td>
<td>Life Sciences</td>
<td>Indiana University School of Medicine/Temple University</td>
</tr>
<tr>
<td>Dryad</td>
<td>Life Sciences</td>
<td>University of Alberta*</td>
</tr>
<tr>
<td>EcoGene</td>
<td>Life Sciences</td>
<td>National Evolutionary Synthesis Center</td>
</tr>
<tr>
<td>eCrystals</td>
<td>Crystallography</td>
<td>University of Southampton</td>
</tr>
<tr>
<td>Ermage</td>
<td>Life Sciences</td>
<td>Medical Research Council</td>
</tr>
<tr>
<td>EMBL</td>
<td>Life Sciences</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>Esther</td>
<td>Life Sciences</td>
<td>Fish National Institute for Agricultural Research*</td>
</tr>
<tr>
<td>Europeat</td>
<td>Social Sciences</td>
<td>European Union</td>
</tr>
<tr>
<td>Finnish Social Science Archive</td>
<td>Social Sciences</td>
<td>University of Tampere</td>
</tr>
<tr>
<td>Gen Expression Omnibus</td>
<td>Genetics</td>
<td>National Center for Biotechnology Information</td>
</tr>
</tbody>
</table>

DCI list of repositories

http://wokinfo.com/products_tools/multidisciplinary/dci/repositories/
What we’ve done and how we’ve done it

1. Data Set Citation (Everyone!)
   - 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.

2. Publication of data sets (Journal publishers)
   - Doi:10232/123

0. Serving of data sets (Data centres)
   - Doi:10232/123ro

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

• 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
• Other data centres/repositories/libraries are also minting DOIs for their data
• 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.

http://www.keepcalm-o-matic.co.uk/default.aspx#createposter
Workflows
Data repository workflows

- Workflows are very varied! No one-size fits all method

- Can have multiple workflows in the same data centre, depending on interactions with external sources (“Engaged submitter”/ “Data dumper” / “Third party requester”)

![Diagram of workflow for BADC/NEODC](image.png)