Making Data a First Class Scientific Output: data citation and publication by NERC’s environmental data centres

Sarah Callaghan, Steve Donegan, Sam Pepler, NCAS British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory
Mark Thorley, Natural Environment Research Council
Nathan Cunningham, Peter Kirsch, Polar Data Centre, British Antarctic Survey
Linda Ault, Patrick Bell, Rod Bowie, National Geoscience Data Centre, British Geological Survey
Adam Leadbetter, Roy Lowry, Gwen Moncoiffe, British Oceanographic Data Centre
Kate Harrison, Ben Smith-Haddon, Anita Weatherby, Dan Wright, Environmental Information Data Centre, Centre for Ecology and Hydrology

sarah.callaghan@stfc.ac.uk @sorcha_ni
... data was hard to capture, but could be (relatively) easily published in image or table format

But now...
there’s simply too much information associated with everything we need to know about a scientific event
- whether that’s an observation, simulation, development of a theory, or any combination of these.

Data always has been the foundation of scientific progress – without it, we can’t test any of our assertions.
“the amount of data generated worldwide...is growing by 58% per year; in 2010 the world generated 1250 billion gigabytes of data”

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

A lot of people are creating a lot of data, and we’re only going to get more of it.

If this is a data deluge – time to start building arks!

Source: IDC Digital Universe Study, sponsored by EMC, May 2010
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.
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.

As part of the NERC Science Information Strategy (SIS) several projects have been created to provide the framework for NERC to work more closely and effectively with its scientific communities in delivering data and information management services.

One of these is the Data Citation and Publication Project
Data Citation and Publication Project

Aims

• To implement publication and citation of datasets held within the NERC data centres.
• To increase NERC’s influence on work to provide and cite data outputs from scientific work in similar ways to scientific papers.
• To demonstrate to the NERC community that data citation and publication is both personally and scientifically advantageous.
• To form partnerships with other organisations with the same goal of data publication to exploit common activities and achieve a wider community buy-in. To this end, project team members are involved with both the SCOR/IODE/MBL WHOI Library Data Publication Working Group, the CODATA-ICSTI Task Group on Data Citation Standards and Practises and the DataCite Working Group on Criteria for Datacentres.

• Provide a reward to scientists who create data for all their efforts in putting their data in one of our data centres.
We draw a clear distinction between:

**publishing/serving** = making available for consumption (e.g. on the web), and

**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

<table>
<thead>
<tr>
<th>Open</th>
<th>Closed (Restricted access and use of data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Scientific peer-reviewed publication Persistent</td>
<td>Scientific peer-reviewed publication Persistent</td>
</tr>
<tr>
<td>Data Centre (through DOIs) Technical review Persistent Mostly open, though embargoes may apply</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Published</th>
<th>Not Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open (e.g. Webpage) No data review No persistence guarantee</td>
<td>Closed (e.g. CD in a drawer) No data review No persistence guarantee</td>
</tr>
</tbody>
</table>
To a scientist, there is little benefit from making their dataset available as a free download from a webpage.

Reputational risk of doing so:
• others might find errors, or
• take advantage of the dataset to earn new research funding

There’s extra work involved in preparing a dataset for use by others.

Data centres are working with scientists to bring data from the closed servers and CDs into an archive where they can be properly curated, with the eventual aim of publication and the dataset author receiving full academic credit for their efforts.
How we’re going to cite (and publish) data

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)
- Pangaea assign DOIs, and ESSD use DOIs to link to the datasets they publish
- The British Library and DataCite gave us an allocation of 500 DOIs to assign to datasets (we got to define what a dataset is).
What sort of data can we/will we cite?

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.

Upgrades to versions of data formats will result in new editions of datasets.
What data centres can do and what we can’t

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.

When we cite (i.e. assign a DOI to) a dataset, we’re confirming that, in our opinion, the dataset meets a level of **technical quality** (metadata and format) and that we will make it available and keep it frozen for the foreseeable future.

The day job – take in data and metadata supplied by scientists (often on a 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.
(Scientific) Communication through the ages

Journals have been the traditional route for disseminating scientific knowledge. Papers work but...

...it’s now becoming more important to ensure that the data that underpin a specific scientific result are available and that the conclusions arising from it can be tested.

If the data’s lost/locked away/stored on obsolete media/in arcane formats/without documentation, how can we do that?

Technology has given us new tools, but it’s also provided new challenges

http://www.intoon.com/#68559
Data journals and scientific publication of data

• 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 recognized academic journals to do this. The timescales for this are quite tight, as we want to tie in with the timescales for the next Intergovernmental Panel on Climate Change (IPCC) report

• Data journals already exist:
  • Earth System Science Data (http://earth-system-science-data.net/)
  • Geochemistry, Geophysics, Geosystems (G3 http://www.agu.org/journals/gc/)

DATA: BY THE NUMBERS

- Number of years to get data: 3
- Number of years to interpret data: 2
- Number of years to write about data: 15
- Number of slides to present data: 1

www.phdcomics.com
Conclusions

• The NERC data citation and publication project has been running for 1 year.
  • We’re in phase 2 of the project (which will take 2 years)
  • At the end of this phase, all the NERC data centres will have:
    • At least 1 dataset with associated DOI
    • 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
  • Our users are already expressing an interest in data citation.

“We share because we do science, not alchemy.”
Jason Priem (Datacite Summer meeting, August 2011)
Step XX: Get occasionally side-tracked searching for data specific comics on the internet
Censormatic picture from: