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 Enviroment 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

Historically speaking...

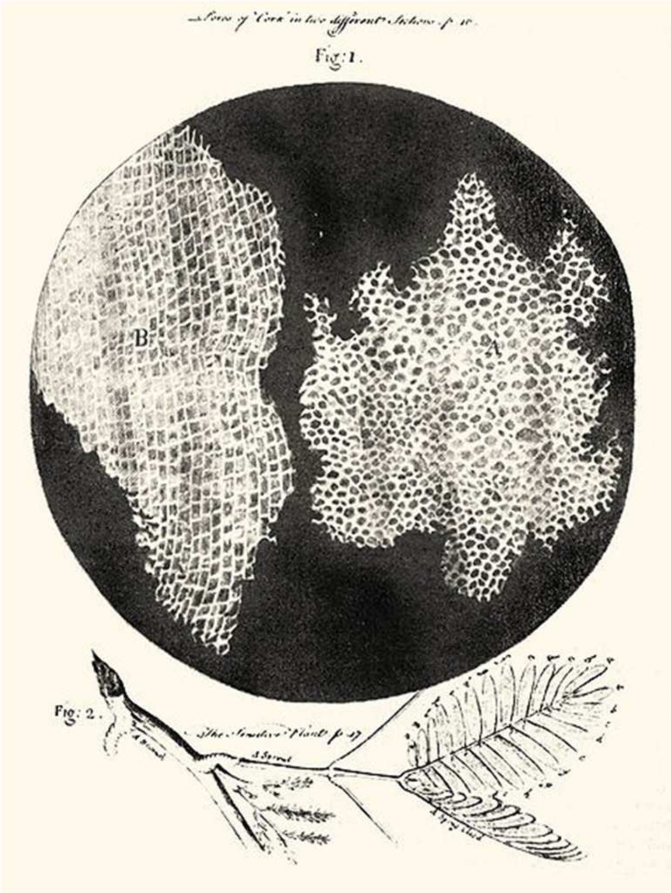
... data was hard to capture, but could be (relatively) easily published in image or table format

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.



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

The Data Deluge

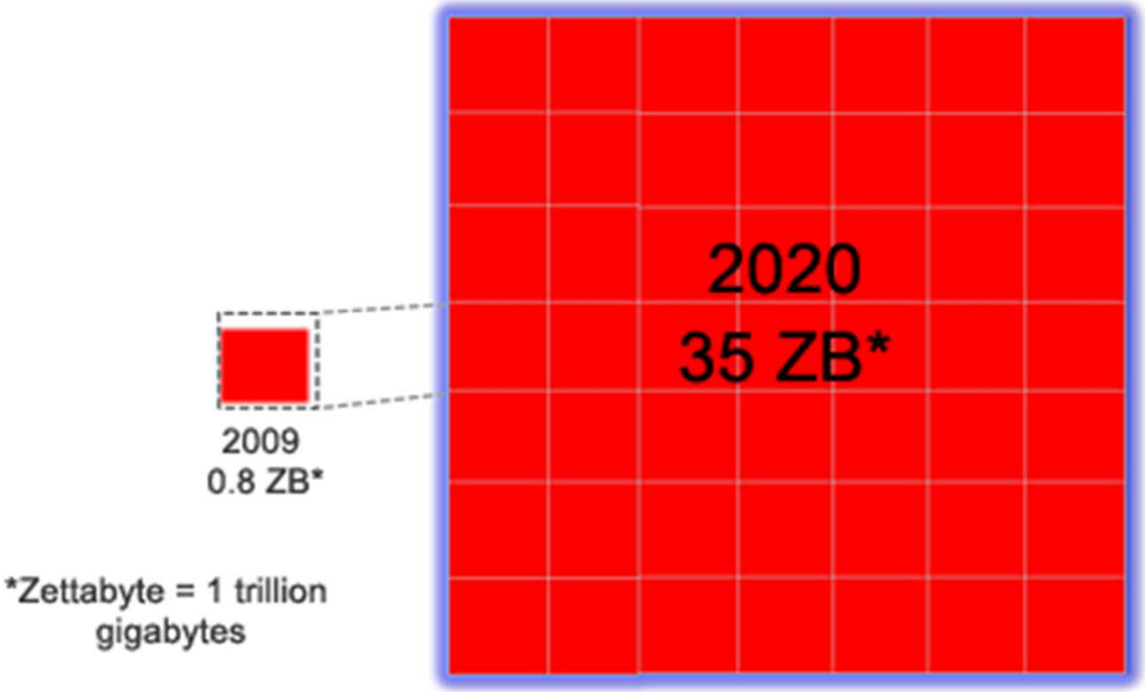
“the amount of data generated worldwide...is growing by 58% per year; in 2010 the world generated 1250 billion gigabytes of data”

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!

Figure 1: The Digital Universe 2009 – 2020
Growing by a Factor of 44



*Zettabyte = 1 trillion gigabytes

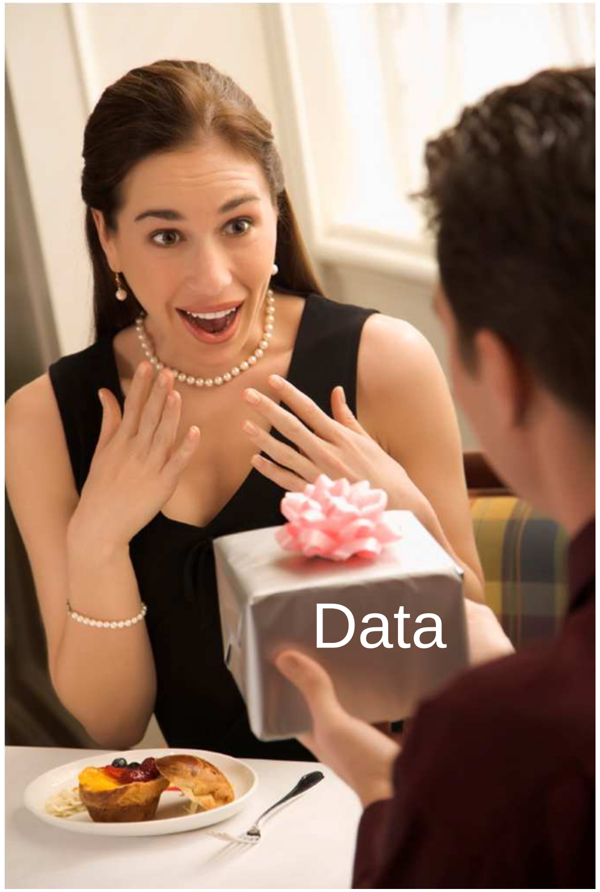
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

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...



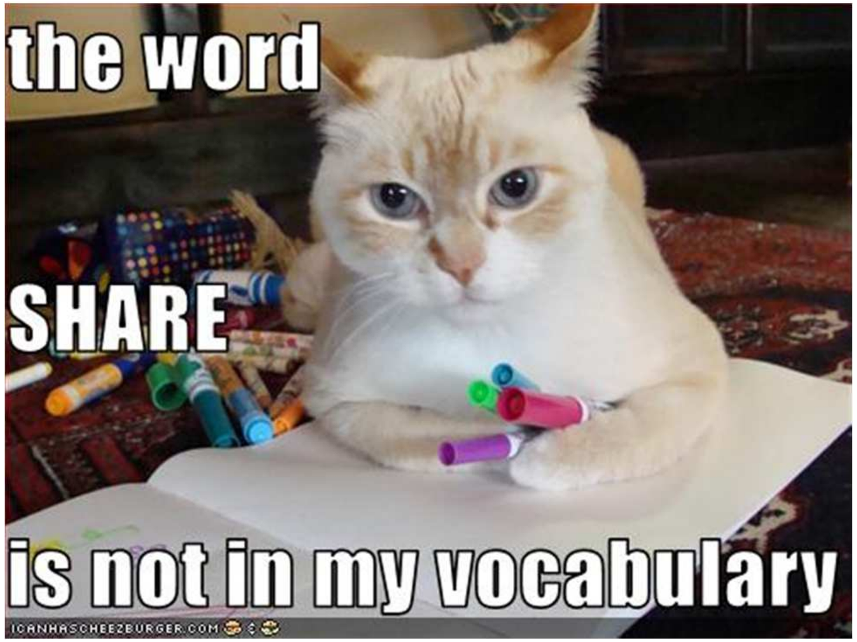
Knowledge is power!

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

British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

NGDC
National Geoscience Data Centre

BODC

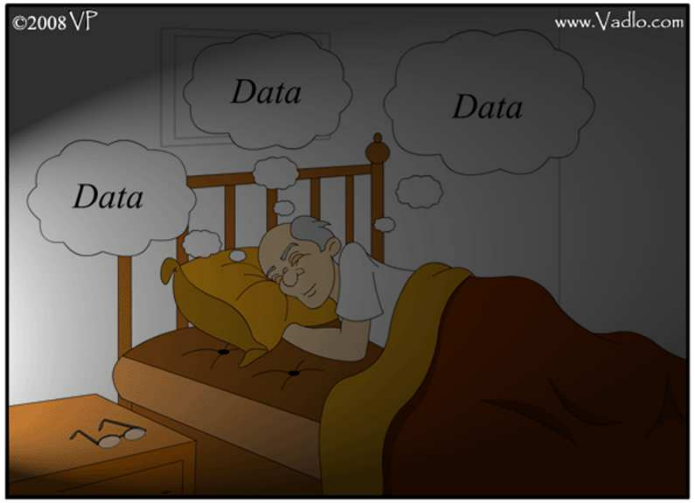
Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NEODC

British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

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.



The Not-So-Secret life of a PI.

- Provide a reward to scientists who create data for all their efforts in putting their data in one of our data centres.

“Publishing” versus “publishing” and “Open” versus “Closed”

We draw a clear distinction between:

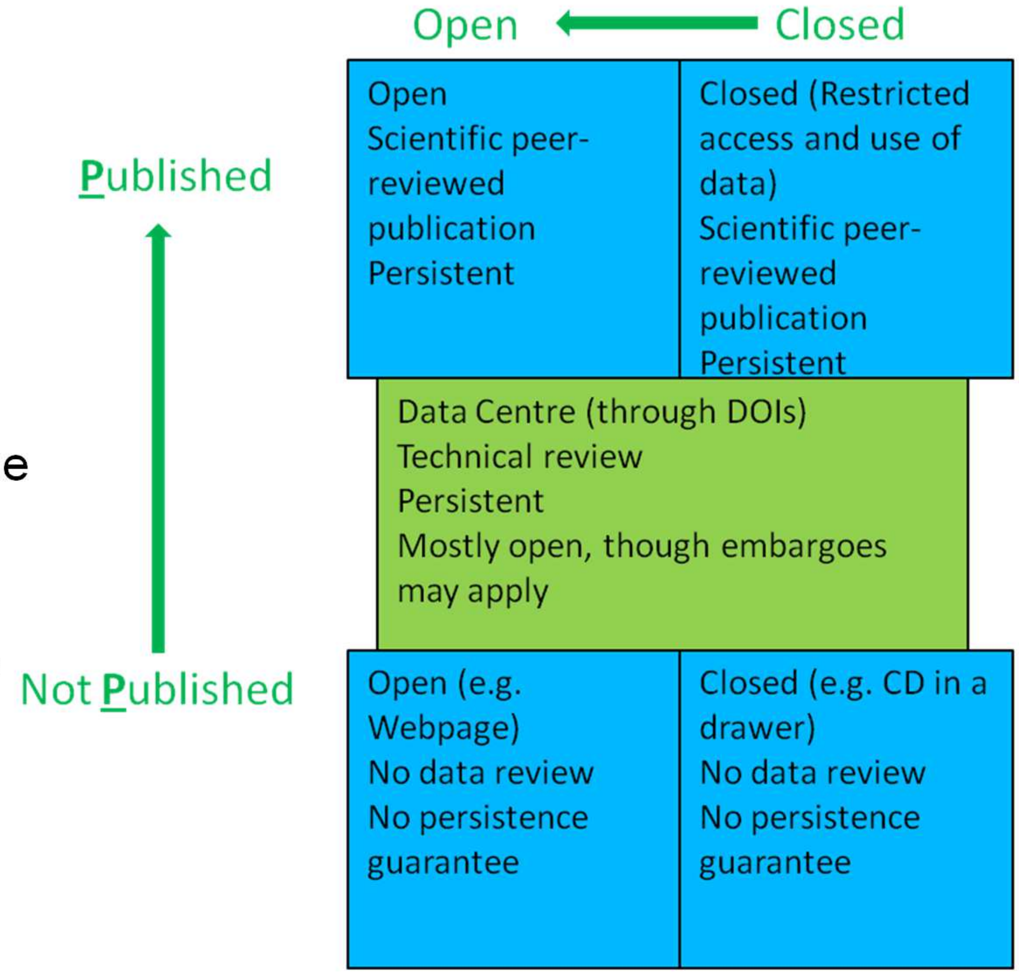
publishing/serving = making available for consumption (e.g. on the web), and

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



Open
Scientific peer-reviewed publication
Persistent

Closed (Restricted access and use of data)
Scientific peer-reviewed publication
Persistent

Data Centre (through DOIs)
Technical review
Persistent
Mostly open, though embargoes may apply

Open (e.g. Webpage)
No data review
No persistence guarantee

Closed (e.g. CD in a drawer)
No data review
No persistence guarantee

“publishing” on the web

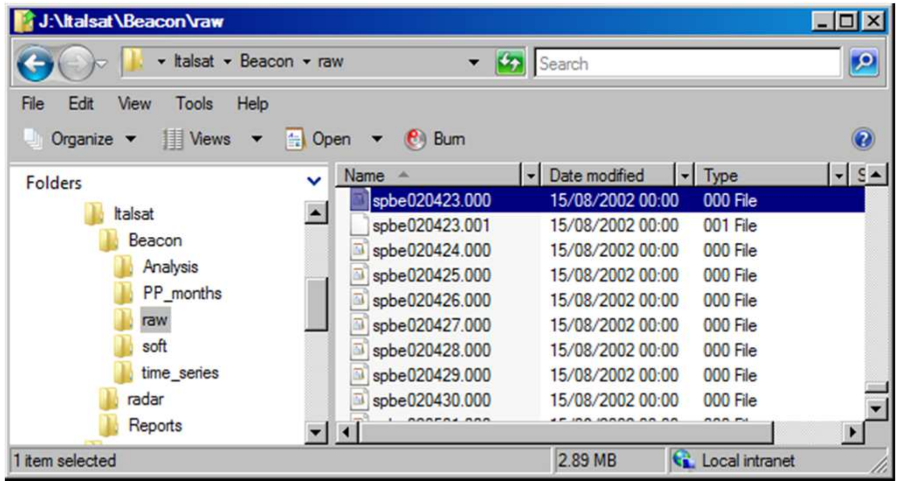
To a scientist, there is little benefit from making their dataset available as a free download from a webpage.

Reputational risk of doing so:

- others might find errors, or
- take advantage of the dataset to earn new research funding

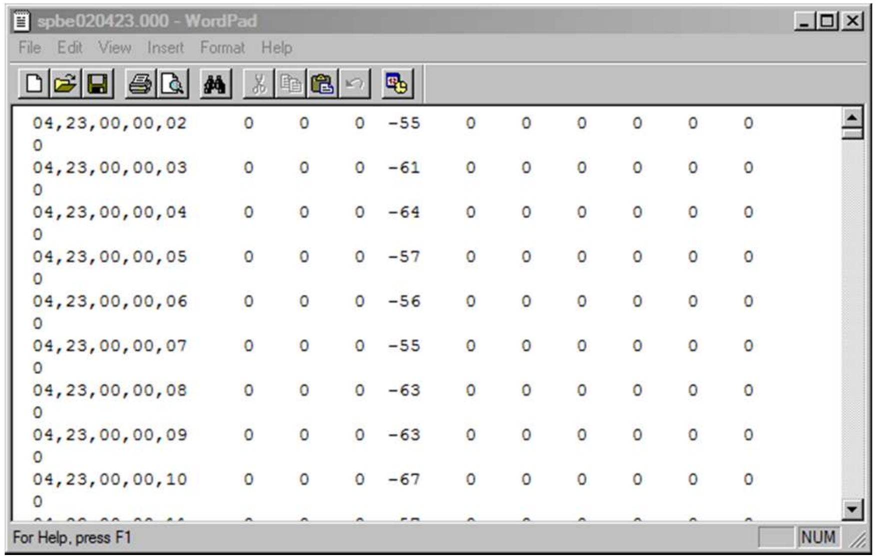
There’s extra work involved in preparing a dataset for use by others.

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.



Name	Date modified	Type
spbe020423.000	15/08/2002 00:00	000 File
spbe020423.001	15/08/2002 00:00	001 File
spbe020424.000	15/08/2002 00:00	000 File
spbe020425.000	15/08/2002 00:00	000 File
spbe020426.000	15/08/2002 00:00	000 File
spbe020427.000	15/08/2002 00:00	000 File
spbe020428.000	15/08/2002 00:00	000 File
spbe020429.000	15/08/2002 00:00	000 File
spbe020430.000	15/08/2002 00:00	000 File

1 item selected | 2.89 MB | Local intranet



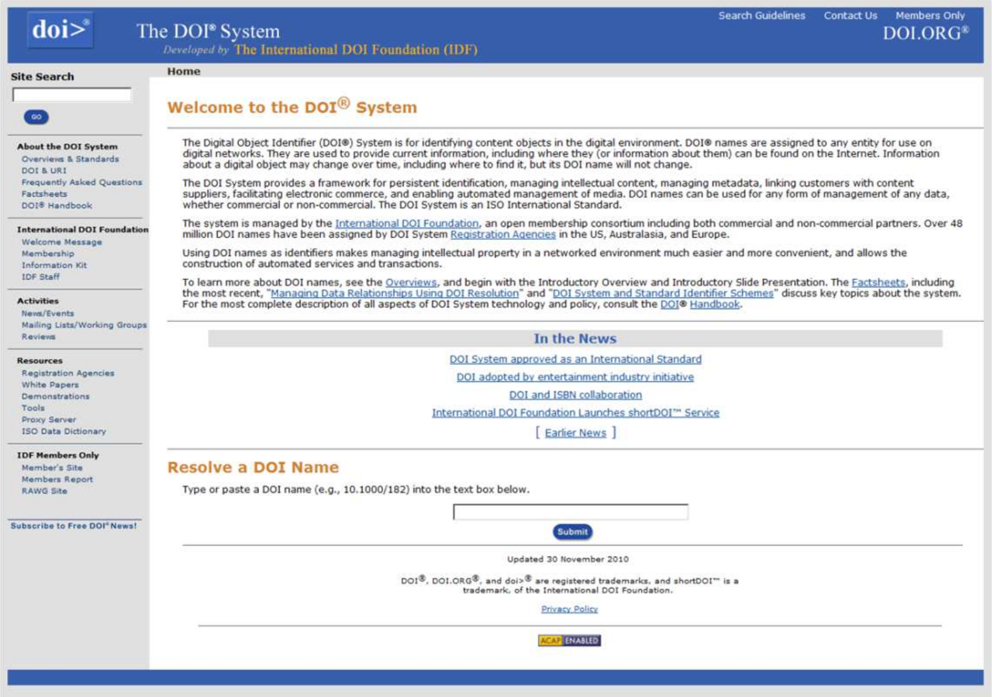
```
04,23,00,00,02     0     0     0  -55     0     0     0     0     0     0
0
04,23,00,00,03     0     0     0  -61     0     0     0     0     0     0
0
04,23,00,00,04     0     0     0  -64     0     0     0     0     0     0
0
04,23,00,00,05     0     0     0  -57     0     0     0     0     0     0
0
04,23,00,00,06     0     0     0  -56     0     0     0     0     0     0
0
04,23,00,00,07     0     0     0  -55     0     0     0     0     0     0
0
04,23,00,00,08     0     0     0  -63     0     0     0     0     0     0
0
04,23,00,00,09     0     0     0  -63     0     0     0     0     0     0
0
04,23,00,00,10     0     0     0  -67     0     0     0     0     0     0
0
```

For Help, press F1 | NUM

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).



Welcome to the DOI® System

The Digital Object Identifier (DOI®) System is for identifying content objects in the digital environment. DOI® names are assigned to any entity for use on digital networks. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI name will not change.

The DOI System provides a framework for persistent identification, managing intellectual content, managing metadata, linking customers with content suppliers, facilitating electronic commerce, and enabling automated management of media. DOI names can be used for any form of management of any data, whether commercial or non-commercial. The DOI System is an ISO International Standard.

The system is managed by the International DOI Foundation, an open membership consortium including both commercial and non-commercial partners. Over 48 million DOI names have been assigned by DOI System Registration Agencies in the US, Australasia, and Europe.

Using DOI names as identifiers makes managing intellectual property in a networked environment much easier and more convenient, and allows the construction of automated services and transactions.

To learn more about DOI names, see the Overviews, and begin with the Introductory Overview and Introductory Slide Presentation. The Factsheets, including the most recent, "Managing Data Relationships Using DOI Resolution" and "DOI System and Standard Identifier Schemes" discuss key topics about the system. For the most complete description of all aspects of DOI System technology and policy, consult the DOI® Handbook.

In the News

DOI System approved as an International Standard

DOI adopted by entertainment industry initiative

DOI and ISBN collaboration

International DOI Foundation Launches shortDOI™ Service

[Earlier News]

Resolve a DOI Name

Type or paste a DOI name (e.g., 10.1000/182) into the text box below.

Updated 30 November 2010

DOI®, DOI.ORG®, and doi>® are registered trademarks, and shortDOI™ is a trademark, of the International DOI Foundation.

Privacy Policy

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

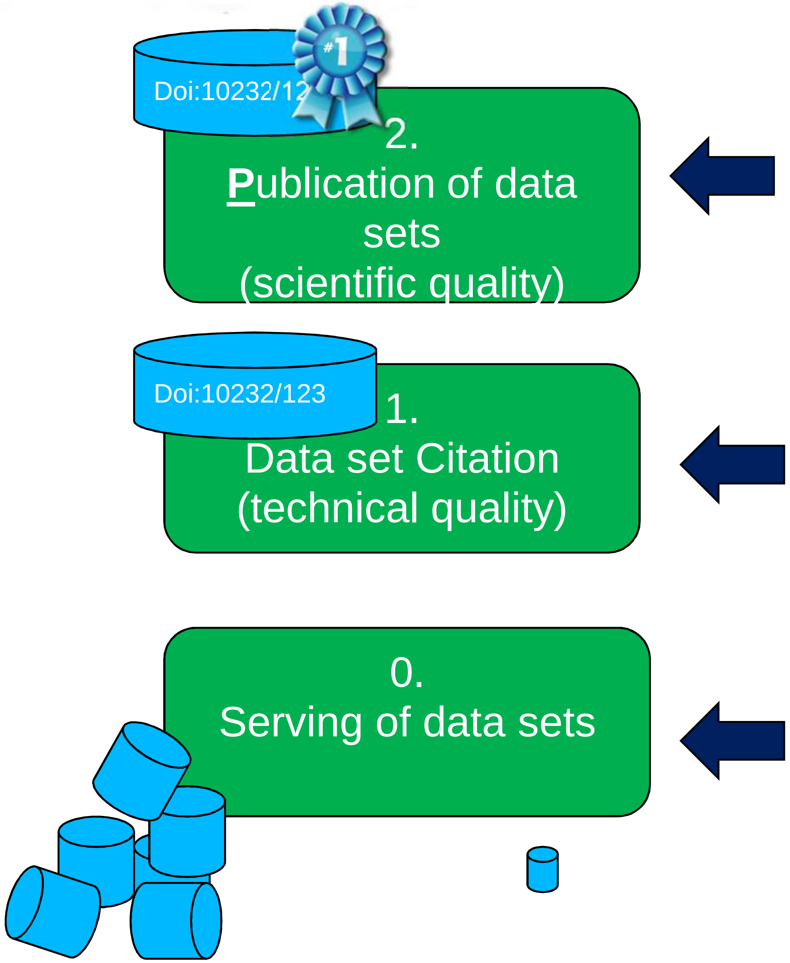
When a dataset is cited that means:

- 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



Doi:10232/12

2.
Publication of data sets
(scientific quality)

Doi:10232/123

1.
Data set Citation
(technical quality)

0.
Serving of data sets

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 forseeable 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.

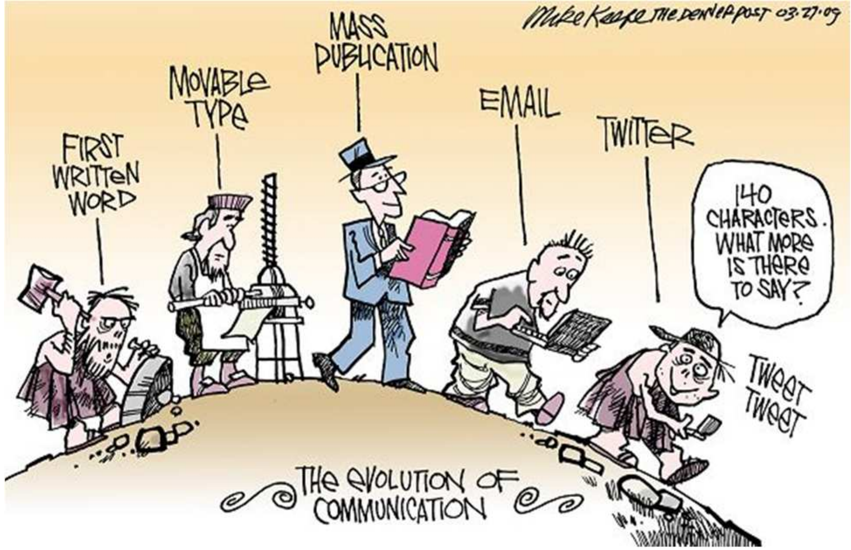
(Scientific) Communication through the ages

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


Mike Keefe THE DENVER POST 03.27.09
FIRST WRITTEN WORD
MOVABLE TYPE
MASS PUBLICATION
EMAIL
TWITTER
140 CHARACTERS. WHAT MORE IS THERE TO SAY?
TWEET TWEET
THE EVOLUTION OF COMMUNICATION

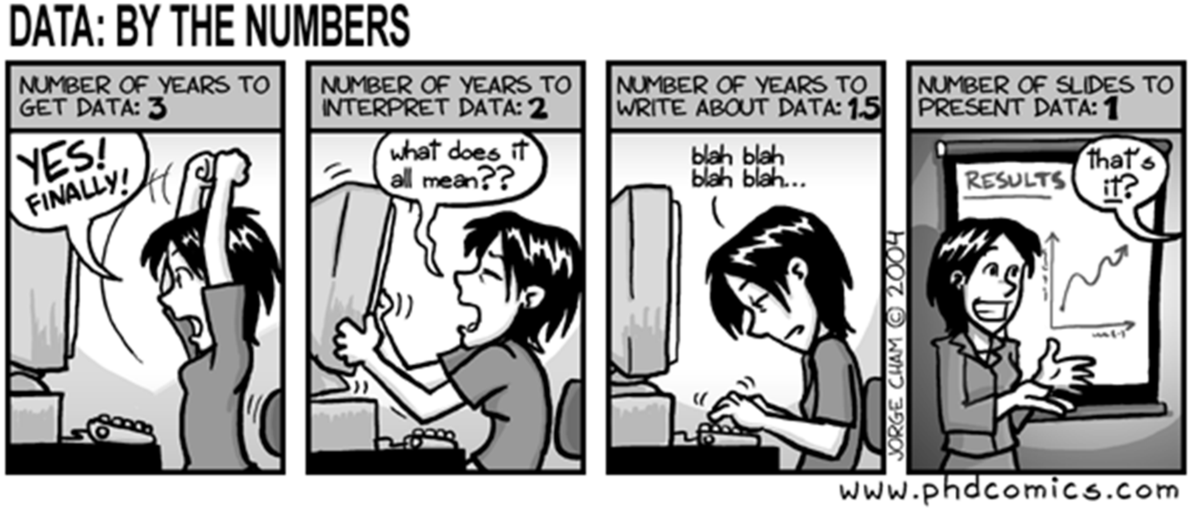
http://www.intoon.com/#68559

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.

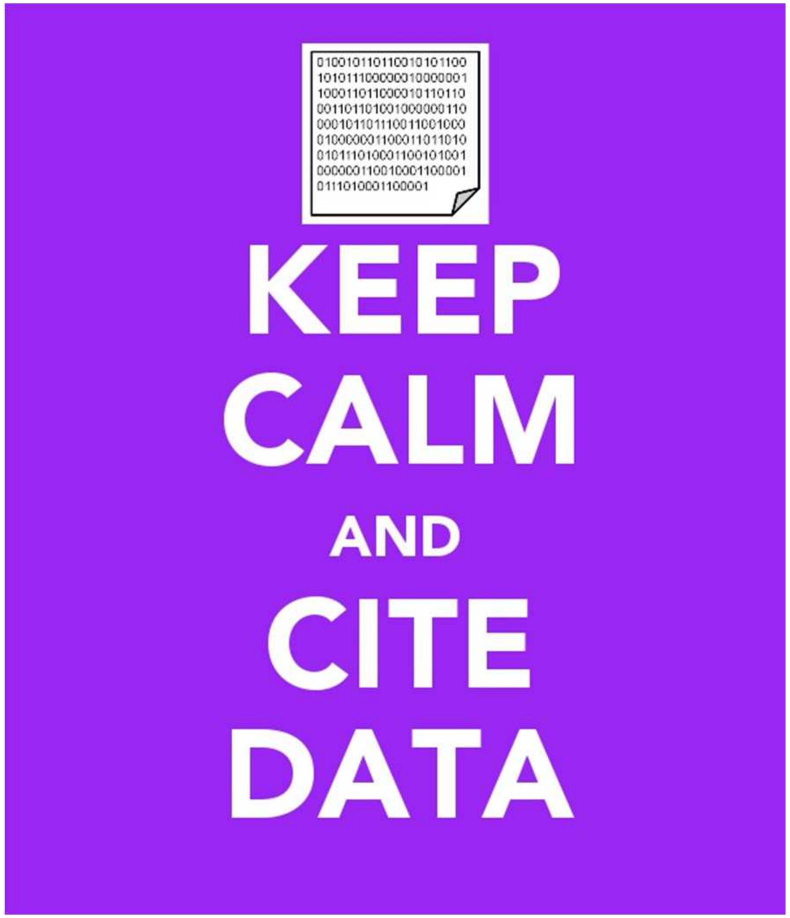
- Publication – and scientific peer-review – is the next step
- We are working with recognized academic journals to do this. The timescales for this are quite tight, as we want to tie in with the timescales for the next Intergovernmental Panel on Climate Change (IPCC) report

- 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
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
JORGE CHAM © 2004
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.

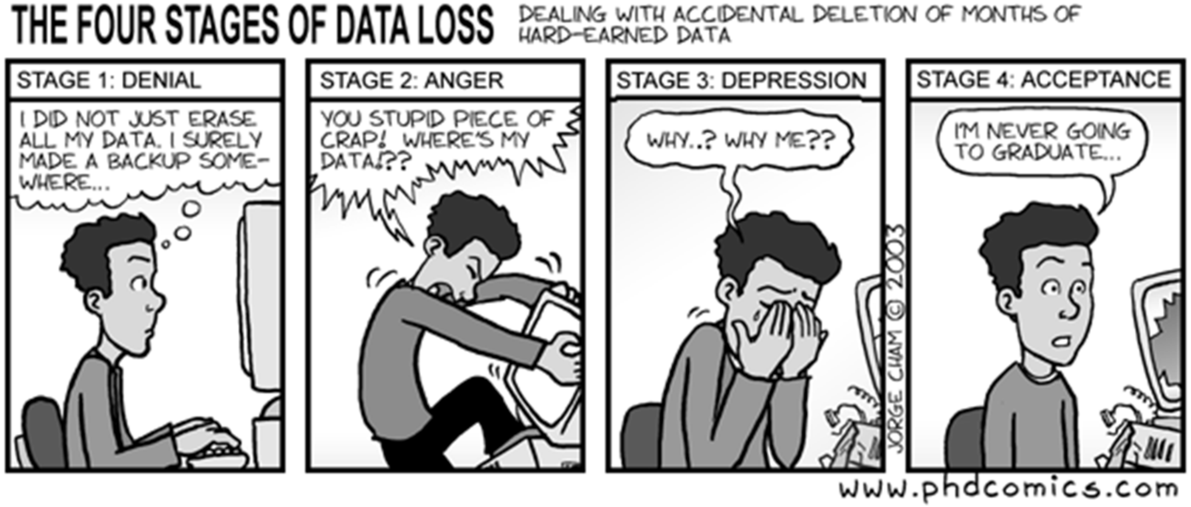


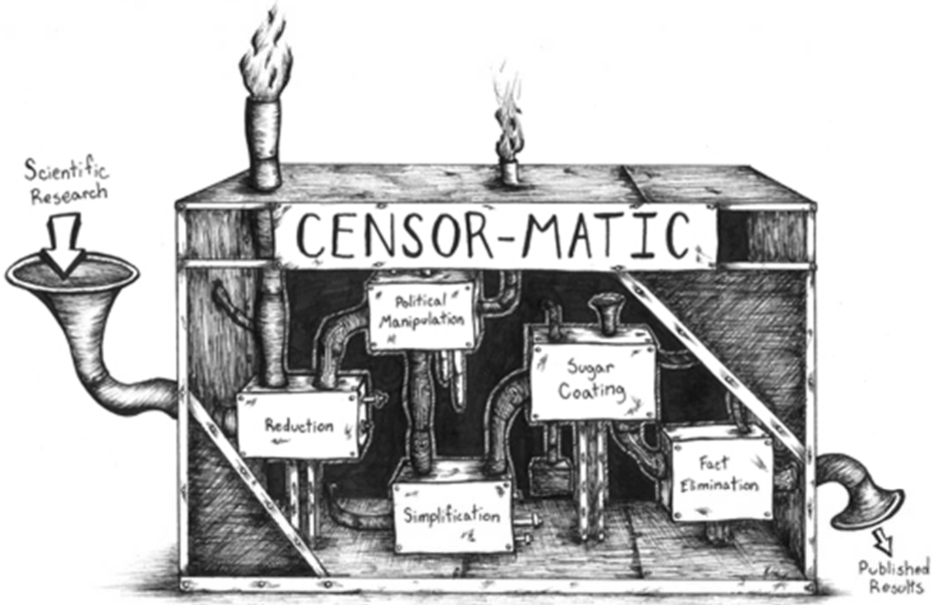
http://www.keepcalm-o-matic.co.uk/default.aspx#createposter

"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


THE FOUR STAGES OF DATA LOSS
DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA
STAGE 1: DENIAL
I DID NOT JUST ERASE ALL MY DATA. I SURELY MADE A BACKUP SOMEWHERE...
STAGE 2: ANGER
YOU STUPID PIECE OF CRAP! WHERE'S MY DATA!??
STAGE 3: DEPRESSION
WHY..? WHY ME??
STAGE 4: ACCEPTANCE
I'M NEVER GOING TO GRADUATE...
JORGE CHAM © 2003
www.phdcomics.com



Censormatic picture from: http://scienceblogs.com/clock/2007/04/framing_politics_based_on_scie_1.php