Science & Technology Facilities Council

DOIs for Data: Progress in Data Citation and Publication in the Geosciences

Sarah Callaghan, Fiona Murphy, Jonathan Tedds, Rob Allan

[sarah.callaghan@stfc.ac.uk]
@sorcha.ni

IN22A. IN22A. Data Stewardship, Citation With Confidence, and Preparing Next Generation of Data Managers

Science & Technology Facilities Council

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

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NEODC

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

Science & Technology Facilities Council

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

Science & Technology Facilities Council

Geoscience Data Journal

Wiley-Blackwell and the Royal Meteorological Society

- Partnership formed between **Royal Meteorological Society** & academic publishers **Wiley-Blackwell**
  - develop a mechanism for the formal publication of data in the Open Access *Geoscience Data Journal (GDJ)*

• 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.

RMetS
Royal Meteorological Society

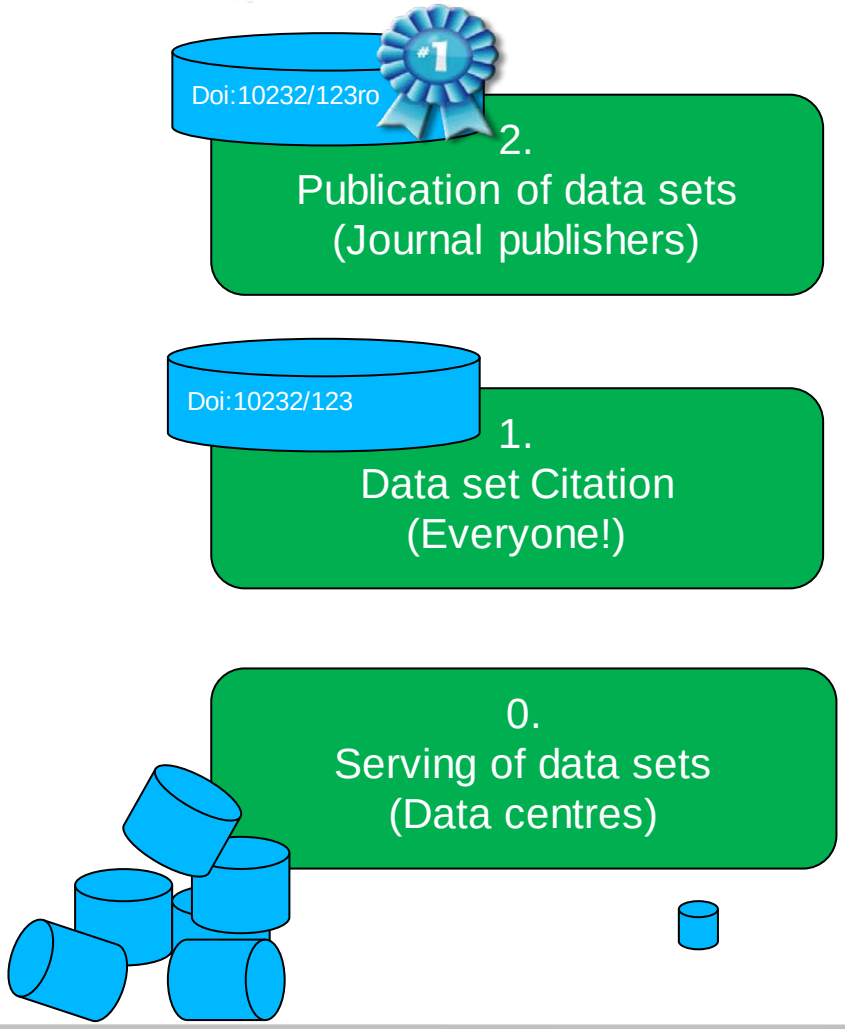
WILEY-BLACKWELL

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

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Identifiers for data and how data centres use them



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 identifiers that are permanent and unambiguous. 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 need identifiers to locate and identify the data in our archive. Note that the data can and does change!

Inside the BADC archive

Home | My BADC | Data | Search | Community | Help

Get Data | Access Rules | Submit Data | Dataset Index

Get Data

Logout Help

Username: scallagh

Download multiple files How to use * Depth: 1 GO!

Current directory: / badc

CDs	*Copies of datasets distributed via CDROM*
abacus	*Arctic Biosphere-Atmosphere Coupling at multiple Scales (ABACUS)*
accmip	*Atmospheric Chemistry & Climate Model Intercomparison Project (ACCMIP)*
acsoe	*Atmospheric Chemistry Studies in the Oceanic Environment (ACSOE)*
active	*Aerosol and Chemical Transport in Tropical Convection (ACTIVE)*
adriex	*Aerosol Direct Radiative Impact Experiment (ADRIEX)*
africa-lam	*Met Office - Limited Area Model for Africa (Africa-LAM) data*
amma	*African Monsoon Multidisciplinary Analysis (AMMA)*
appraise	*Aerosol Properties, Processess and Influences on the Earth's Climate (APPRAISE) Directed Mode Programme*
armagh	*The Armagh Observatory Climate Data (1838 - present)*
autex-wintex	*Autumn and Winter Experiments (AUTEX / WINTEX)*
bas	*British Antarctic Survey Data*
bodeker-scientific-ozone	*Bodeker Scientific global vertically resolved Ozone database*
bolton	*The Bolton Experiment (1999-2002)*
bortas	*BORTAS: Quantifying the impact of BOReal forest fires on T*
caesar	*CAESAR - Cirrus and Anvils: European Satellite and Airborr*
capeverde	*Cape Verde Atmospheric Observatory*
cascade	*CASCADE - Scale interactions in the tropical atmosphere*
ccmval	*CCMVal Data*
chablis	*Chemistry of the Antarctic Boundary Layer and the Interface*
chilbolton	*Chilbolton Facility for Atmospheric and Radio Research*
cira	*Cospar International Reference Atmosphere (CIRA 86)*
clace	*CLoud, Aerosol Characterisation Experiment (CLACE) in the*

Datasets can and do change as files get added/changed or moved around the archive.

Dumping our users straight into a list of files isn't the friendliest thing to do…

Home | My BADC | Data | Search | Community | Help

Get Data | Access Rules | Submit Data | Dataset Index

Get Data

Logout Help

Username: scallagh

Download multiple files How to use * Depth: 1 GO!

Current directory: / badc / chilbolton / data / gbs_sparsholt / 2005 / 03

Dataset: *Measurements from the Chilbolton Facility for Atmospheric and Radio Research (CFARR)* Details

Sparsholt Global Broadcasting System receiver - Chilbolton dataset

00README	315 bytes
cfarr-gbs_sparsholt_20050301.nc	694740 bytes
cfarr-gbs_sparsholt_20050302.nc	694740 bytes
cfarr-gbs_sparsholt_20050303.nc	694740 bytes
cfarr-gbs_sparsholt_20050304.nc	694740 bytes
cfarr-gbs_sparsholt_20050305.nc	694740 bytes
cfarr-gbs_sparsholt_20050306.nc	694740 bytes
cfarr-gbs_sparsholt_20050307.nc	694740 bytes
cfarr-gbs_sparsholt_20050308.nc	694740 bytes
cfarr-gbs_sparsholt_20050309.nc	694740 bytes
cfarr-gbs_sparsholt_20050310.nc	694740 bytes
cfarr-gbs_sparsholt_20050311.nc	694740 bytes
cfarr-gbs_sparsholt_20050312.nc	694740 bytes
cfarr-gbs_sparsholt_20050313.nc	694740 bytes
cfarr-gbs_sparsholt_20050314.nc	694740 bytes
cfarr-gbs_sparsholt_20050315.nc	694740 bytes
cfarr-gbs_sparsholt_20050316.nc	694736 bytes
cfarr-gbs_sparsholt_20050317.nc	694736 bytes
cfarr-gbs_sparsholt_20050318.nc	694736 bytes
cfarr-gbs_sparsholt_20050319.nc	694740 bytes
cfarr-gbs_sparsholt_20050320.nc	694740 bytes

Metadata Catalogue

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Search for [ ] in [All] [Go]

Measurements from the Chilbolton Facility for Atmospheric and Radio Research (CFARR)

General Info

Title: Measurements from the Chilbolton Facility for Atmospheric and Radio Research (CFARR)
Type: Data Entity
Sub-Type: Measurement
Abbreviation: Chilbolton (CFARR)
Publication State: published

URI: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_chobs

Summary

Data from observations made using Chilbolton Facility for Atmospheric and Radio Research (CFARR).The Science and Technology Facilities Council (STFC) facility at Chilbolton Observatory, Hampshire (51.1445N, 1.4270W) is the home of many observation systems for meteorological and atmospheric science research. There are 4 radar systems designed to study precipitation, clouds and clear air, of which the largest is the 3 GHz Doppler radar (CAMRa) on the 25 m dish. There are also 4 lidar systems providing data on elastic backscattering, Doppler velocity, water vapour profiles and depolarisation. A wide range of meteorological and multiple raingauge data are available from both Chilbolton and the nearby Sparsholt field site. There is a wide range of radiometers at the site: microwave (for water vapour and liquid water measurements) and downwelling infra-red and visible detectors for radiation budget measurements. This dataset holds attenuation time-series data from vertically polarised 5 km links from South Wonston to Sparsholt. Cloud camera data from the Chilbolton site are available to provide visual information on weather conditions.

CFARR is funded by the Natural Environment Research Council (NERC) and is owned and operated by the Space Science and Technology Department of the STFC.

Content

Introduction

The Chilbolton Facility for Atmospheric and Radio Research (CFARR) is a ground-based atmospheric remote sensing facility in the village of Chilbolton near Winchester in Hampshire

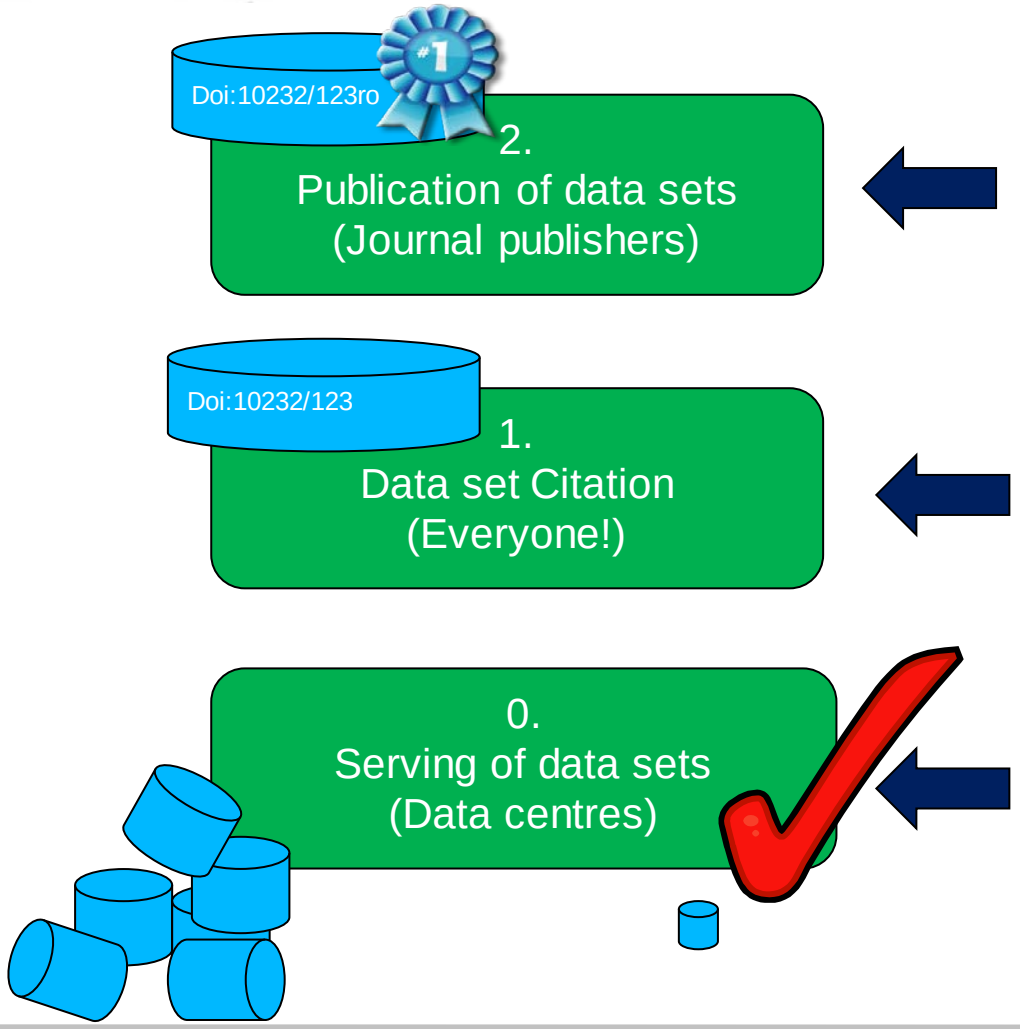
Provides supporting information about the data so that the user can:

- Be confident they've found what they were looking for in the first place
- Know how to open and read the data files
- Cite the data
- Find the data again
- Search for other data

Citation

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Wrench, C.L.]. Chilbolton Facility for Atmospheric and Radio Research (CFARR) data, [Internet]. NCAS British Atmospheric Data Centre, 2003-,*Date of citation*. Available from http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_chobs.

Identifiers for data (2)



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 identifiers that are permanent and unambiguous. 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

URIs, URNs, GUIDs
Identifiers for the data files and the metadata catalogue pages

Why do we want to cite and publish data?

http://www.evidencebased-management.com/blog/2011/11/04/new-evidence-on-big-bonuses/

- Pressure from the UK government to make all data from publicly funded research available to the public for free.
  - Scientists still want to receive attribution and credit for their work
  - General public want to know what the scientists are doing (Climategate...)

- Research funders want reassurance that they're getting value for money from their funding
  - Relies on peer-review of science publications (well established) and data (not done yet!)

- Allows the wider research community to find and use datasets outside their immediate domain, confident that the data is of reasonable quality

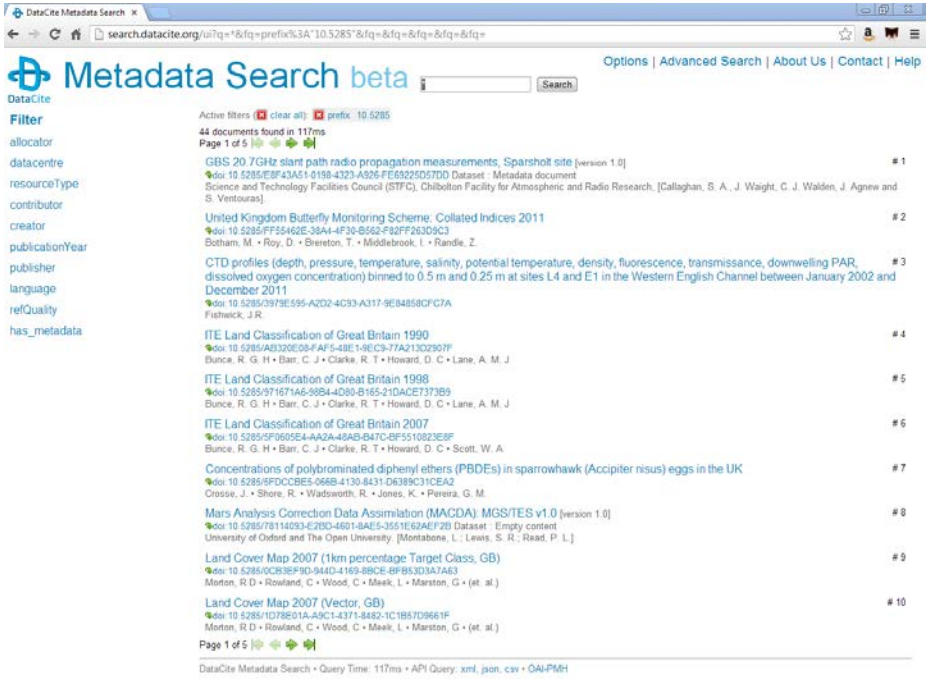
- From a strict data-centric point of view, citation and publication provides an extra incentive for scientists to submit their data to us in appropriate formats and with full metadata!

Science & Technology Facilities Council

How we (formally) cite 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)
- There are moves by academic journal publishers (e.g. Nature) to require data sets to 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 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 the foreseeable future
- 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

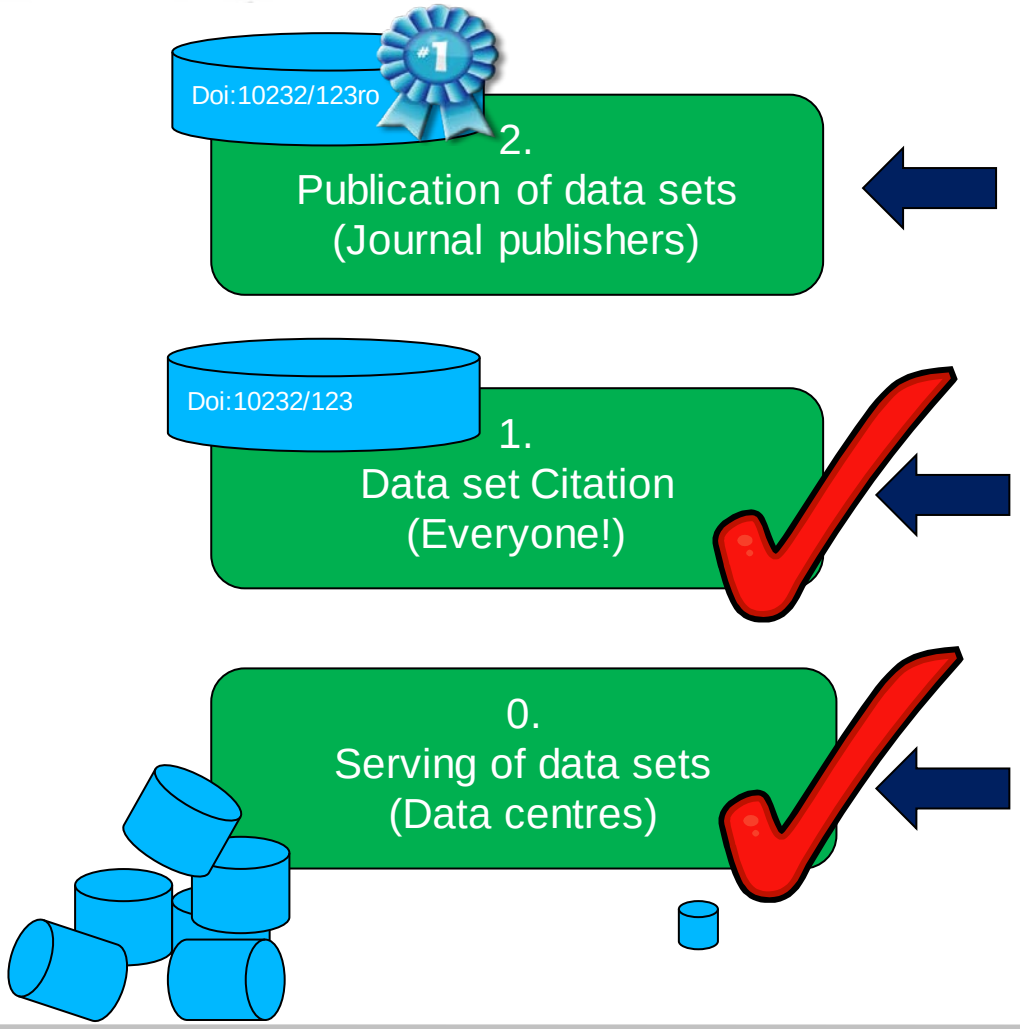
When a dataset is cited that means:

- 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.

Identifiers for data (3)



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.

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.

URIs, URNs, GUIDs
Identifiers for the data files and the metadata catalogue pages

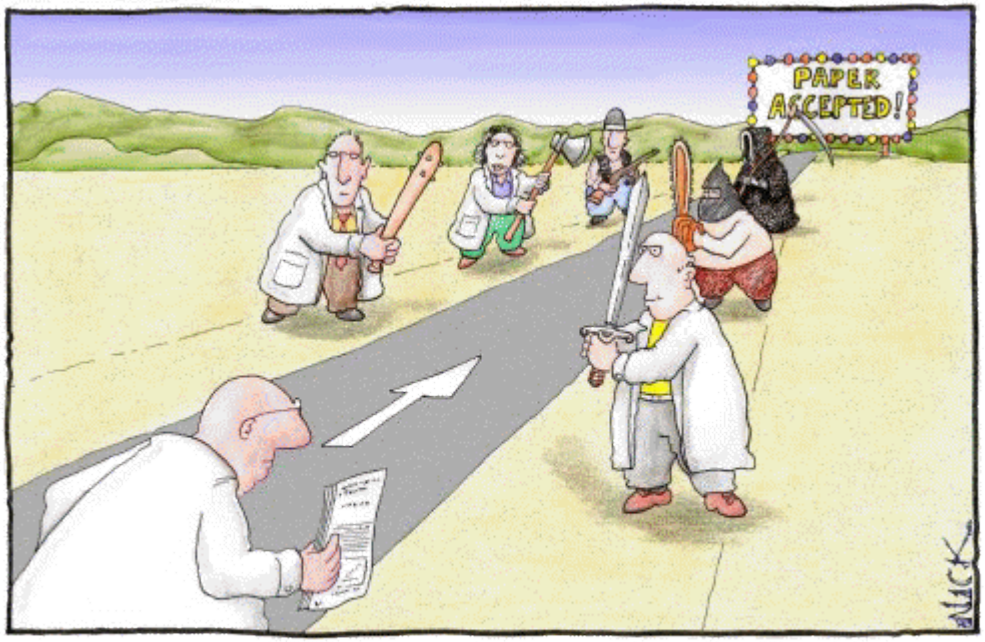
Publishing data for the scholarly record

• 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

"Publishing" versus "publishing" and "Open" versus "Closed"

We draw a clear distinction between:

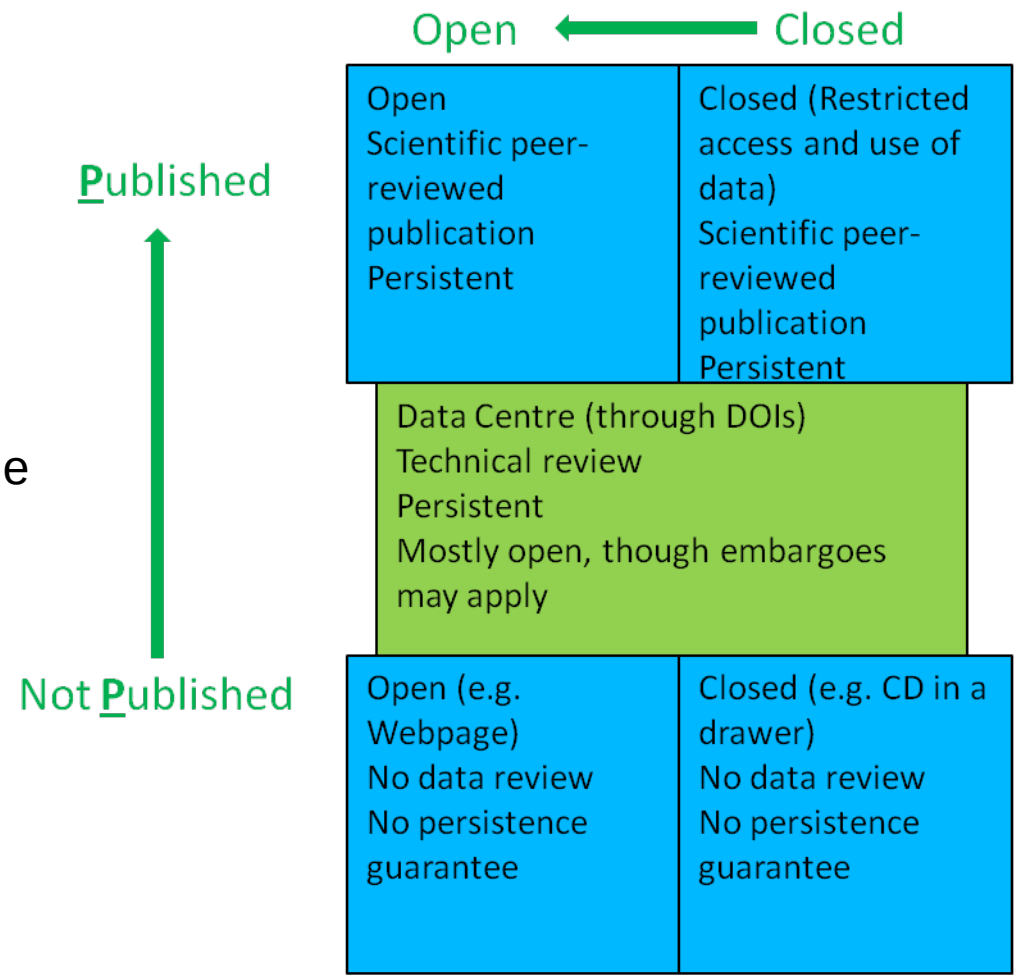
publishing/serving = making available for consumption (e.g. on the web), and

Publishing = publishing after some formal process which adds value for the consumer:

- e.g. PloS ONE type review,or
- more traditional peer review.

AND

- provides commitment to persistence



"publishing" on the web

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

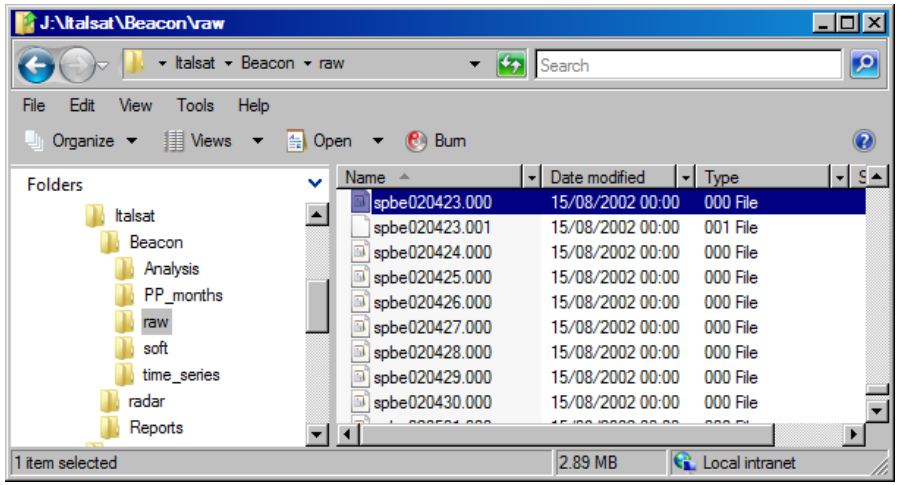
Reputational risk of doing so:

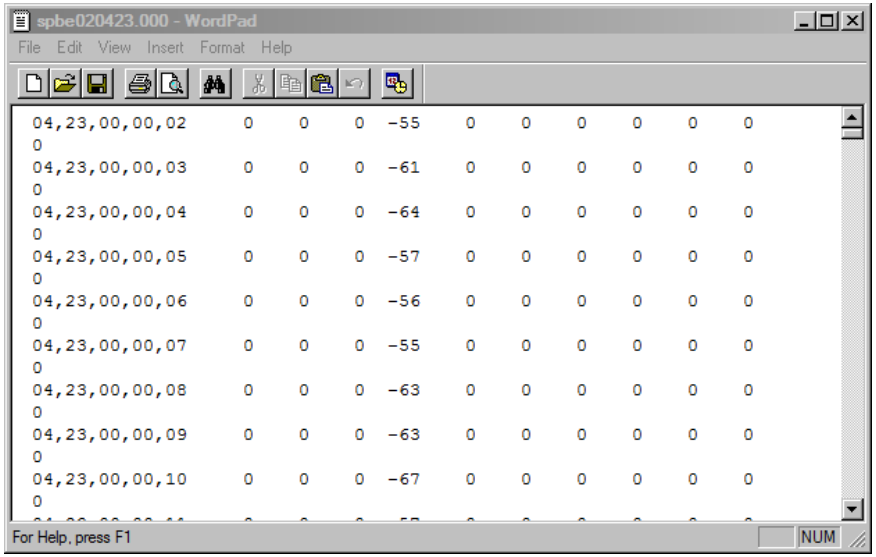
- others might find errors, or
- take advantage of the dataset to earn new research funding

Even when sharing is mandated, there are simple ways of stopping people from using data openly posted on-line (e.g. incomprehensible filenames…)

There's extra effort involved in preparing a dataset for use by others.

Data centres know this extra work is needed, and we want to make sure the dataset author gets credit!





PREPARDE objectives

- capture and manage workflows required to operate the Geoscience Data Journal
  - from submission of a new data paper and dataset, through review and to publication
- develop procedures and policies for authors, reviewers and editors
  - allow the Geoscience Data Journal to accept data papers as submissions for publication
  - focus on guidelines for scientific reviewers who will review the datasets
- incorporate some technical developments at the point of submission
  - data visualisation checks
  - interface improvements
  - enhance the resulting data publications
- put in place procedures needed for data publication in the California Digital Library
- interact with the wider scientific and data community
  - provide recommendations on accreditation requirements for data repositories
- engage the user and stakeholder community
  - promote long-term sustainability and governance of data journals

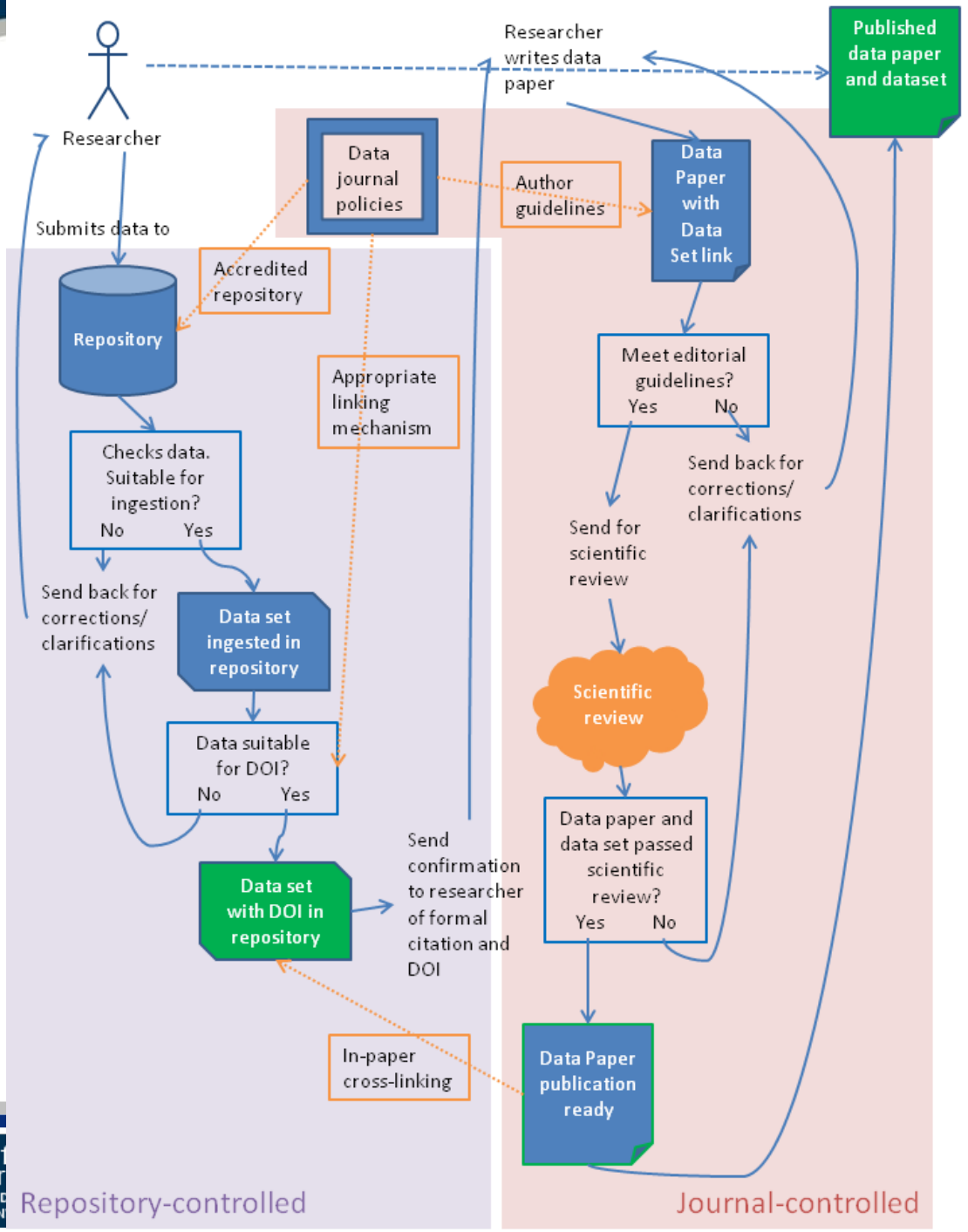
Engraving of printer using the early Gutenberg letter press during the 15th century.
Date unknown - estimate 16th - 19th century
http://commons.wikimedia.org/wiki/File:Gutenberg_press.jpg

Science & Technology Facilities Council

Operating a data journal will be more complicated than a traditional journal, and will require close linking with partnering data repositories.

Figure shows the (potential) steps/workflow required for a researcher to publish a data paper

- Items in orange refer to areas where further work is required and technologies developed.
- Division of area of responsibilities between
  - *repository controlled* processes
  - *journal controlled* processes



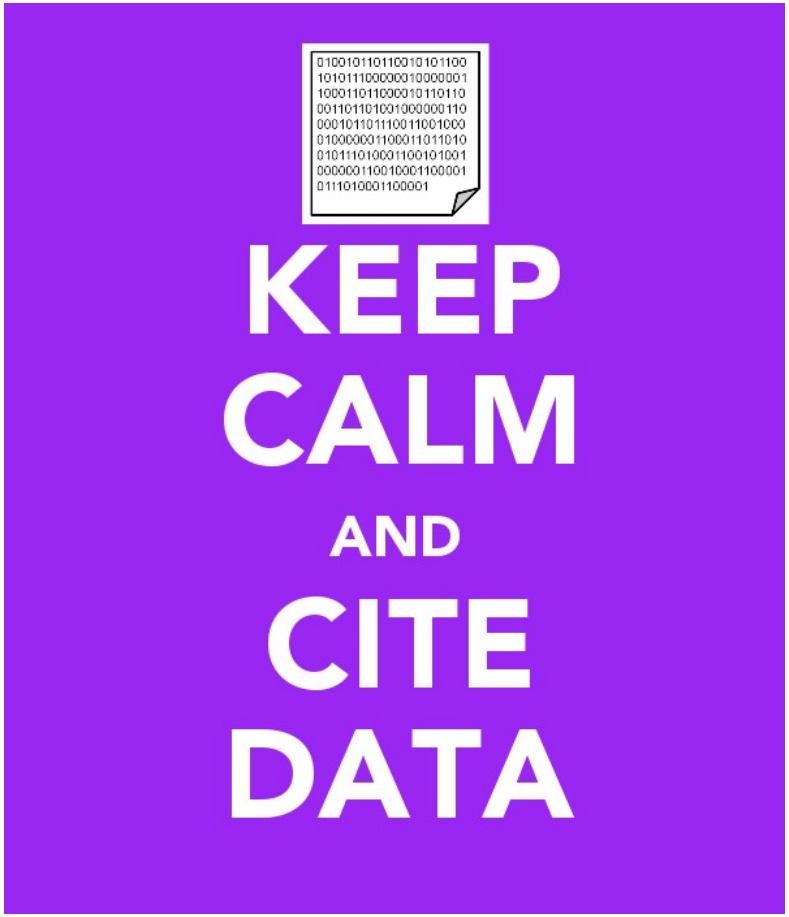
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Data Ar
SCIENCE AND
NATURAL EN

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 that will go into 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 the Geoscience Data Journal), and discussions with Elsevier and Thompson-Reuters.
- The next big step is tackling the thorny issue of peer-review of data – PREPARDE.



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

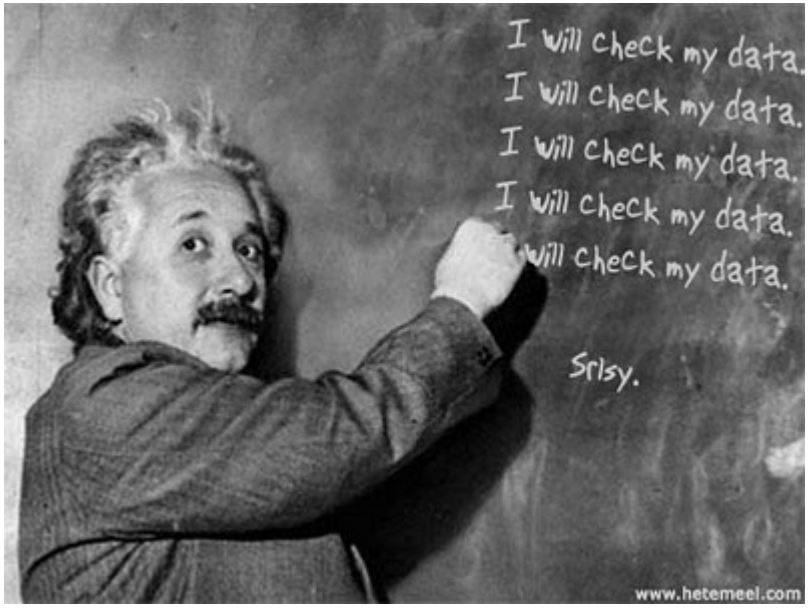


Image credit: Borepatch http://borepatch.blogspot.com/2010/06/its-not-what-you-dont-know-that-hurts.html

WEDNESDAY, DECEMBER 05, 2012

TH32F. TH32F. Publishing Research Data: Peer Review, Data Center Accreditation, and Linking

Convener(s): Fiona Murphy (John Wiley & Sons Ltd) and Sarah Callaghan (STFC)

12:30 PM - 1:30 PM; 2007 (Moscone West)

Thanks!
Any questions?