

Credit where it's due: Data citation and publication in the geosciences

Sarah Callaghan*
sarah.callaghan@stfc.ac.uk
@sorchani

*and many others, including members of the PREPARDE and NERC data citation and publication project teams and the CODATA working group on data citation

ODIN first year conference, 17th October 2013

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



Data, Reproducibility and Science

Science should be reproducible – other people doing the same experiments in the same way should get the same results.

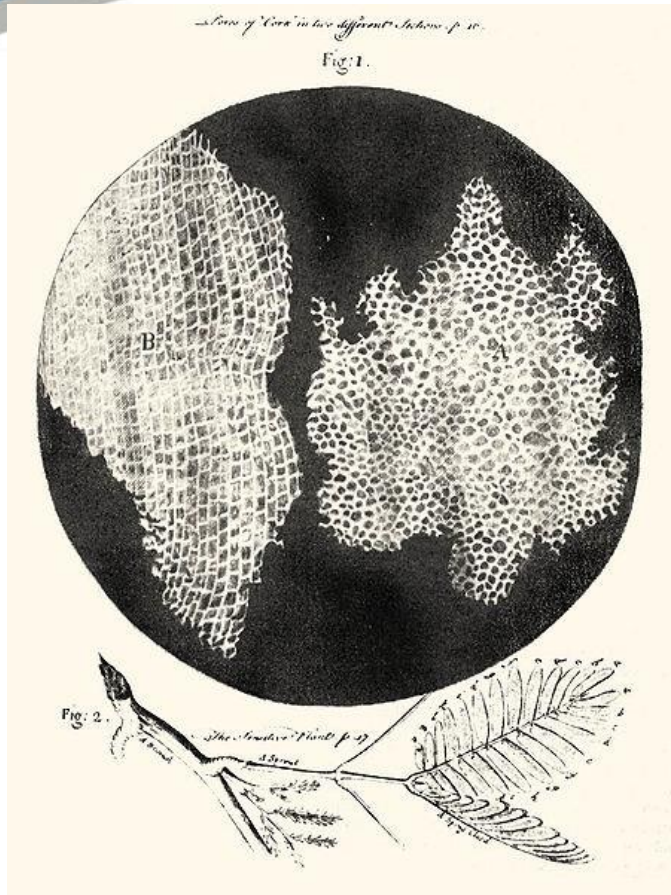
Observational data is not reproducible (unless you have a time machine!)

Therefore we need to have access to the data to confirm the science is valid!



<http://www.flickr.com/photos/31333486@N00/1893012324/sizes/o/in/photostream/>

Journals have always published data...



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

[Observations of Stars in the Spiral Nebula. H. 1622.

The spiral form of this nebula is very distinctly seen in the Pulkova 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.

Observations.

Date.	Object.	Magnitude.	Ang. Pos.	No. of measures.	Distance.	No. of measures.
1851, April 7.	N n	14 55'	5	267.1	4
	N a	a = (11)	229 24	3	88.0	3
	N b	b = (11.12)	109 12	3	242.6	3
	a b	93 42	3	298.6	3
April 28.	a b	94 23	3	300.8	4
	N a	228 36	4		
	N b	108 54	4		
	n a	283 42	3		
	n b	153 30	3		
	a d	d = (12.13)	323 51	3		
	N d	277 27	3		
	a e	e = (13)	112 13	3		
	N e	161 56	3		
	N f	f = (12.13)	309 18	3		
	a f	237 31	3		
May 3.	a g	g = (12.13)	335 23	3	115.5	4
	a h	h = (12.13)	215 17	3		
	g h	193 29	3		
	g k	87 5	3		
	N k	k = (13.14)	51 47	3		
	n k	173 29	4		
	b k	317 23	3		
	b l	l = (11.12)	27 20	4		
	n l	83 17	4	335.2	4
	a e	112 56	4		
	N e	161 39	3		
	a m	m = (12.13)	172 43	5		
	N m	190 44	4		
	b m	238 50	4		
	N a	229 12	4	87.0	3
	N n	14 47	4	264.2	3

The Scientific Papers of William Parsons, Third Earl of Rosse 1800-1867

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

Reasons for citing and publishing data

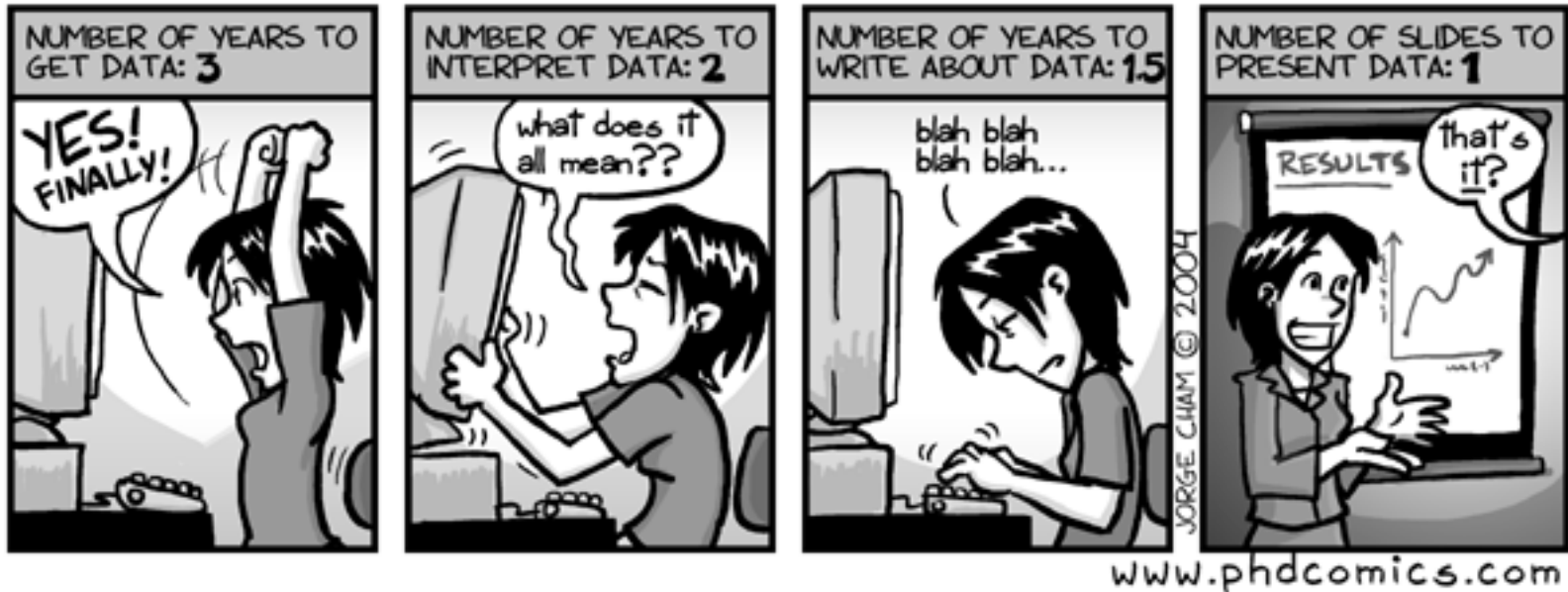
- **Pressure** from (UK) **government** to make data from publicly funded research available for free.
 - **Scientists** want **attribution** and **credit** for their work
 - **Public** want to know what the scientists are doing
 - Good for the **economy** if new industries can be built on scientific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



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

Creating a dataset is hard work!

DATA: BY THE NUMBERS



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

And it takes a long time.

Managing and archiving data so that it's understandable by other researchers is difficult and time consuming too.

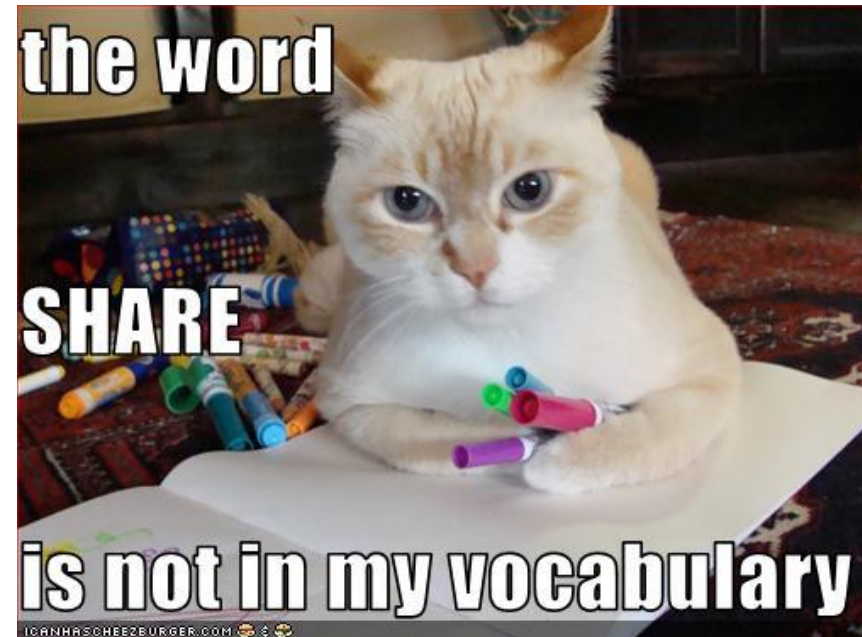
Knowledge is power!

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.

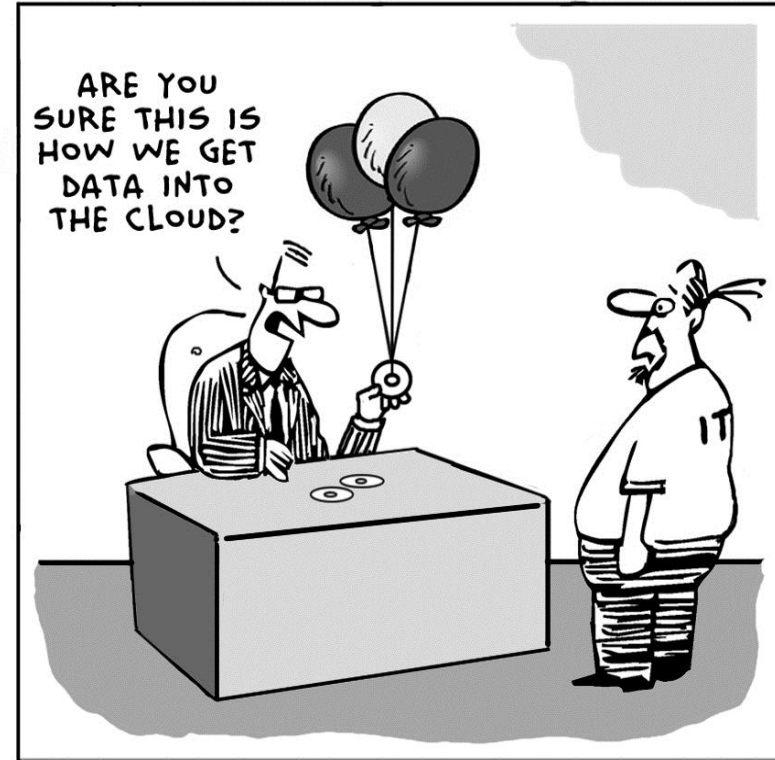
This behaviour comes at a cost for the wider scientific community.



**But if we publish the data,
precedence is established and
credit is given!**

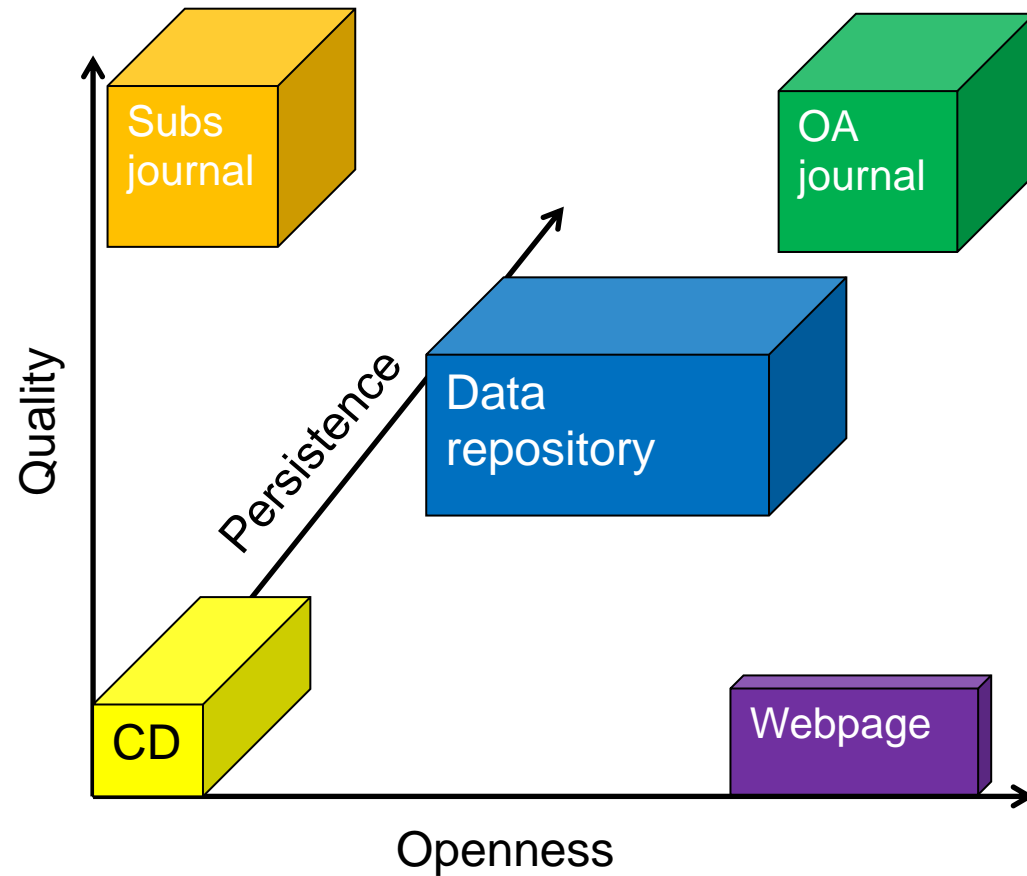
How to publish data

- 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
<http://www.cloudtweaks.com/2011/05/the-lighter-side-of-the-cloud-data-transfer/>

Open/Closed/Published/unpublished



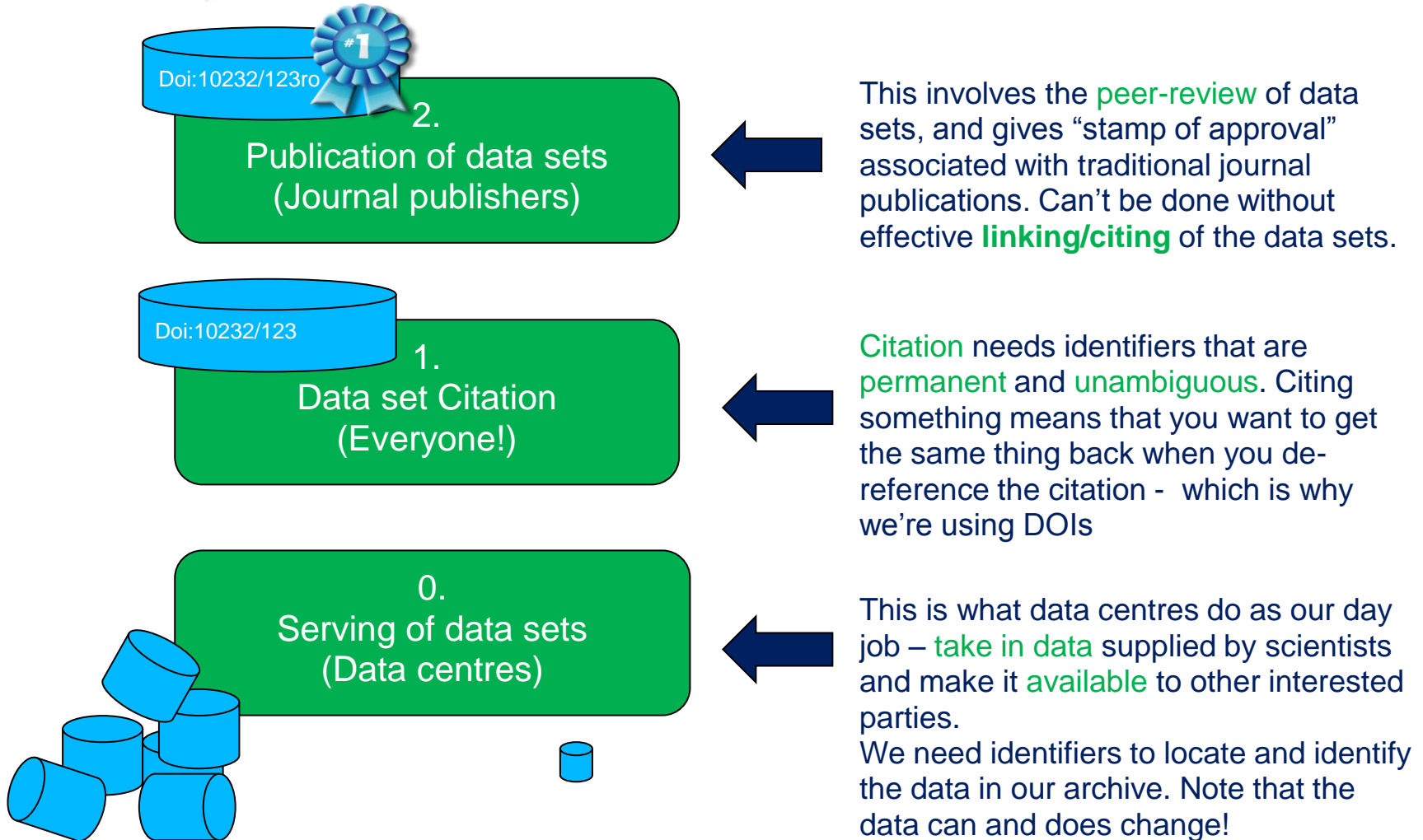
We want to encourage researchers to make their data:

- Open
- Persistent
- Quality assured:
 - through scientific peer review
 - or repository-managed processes

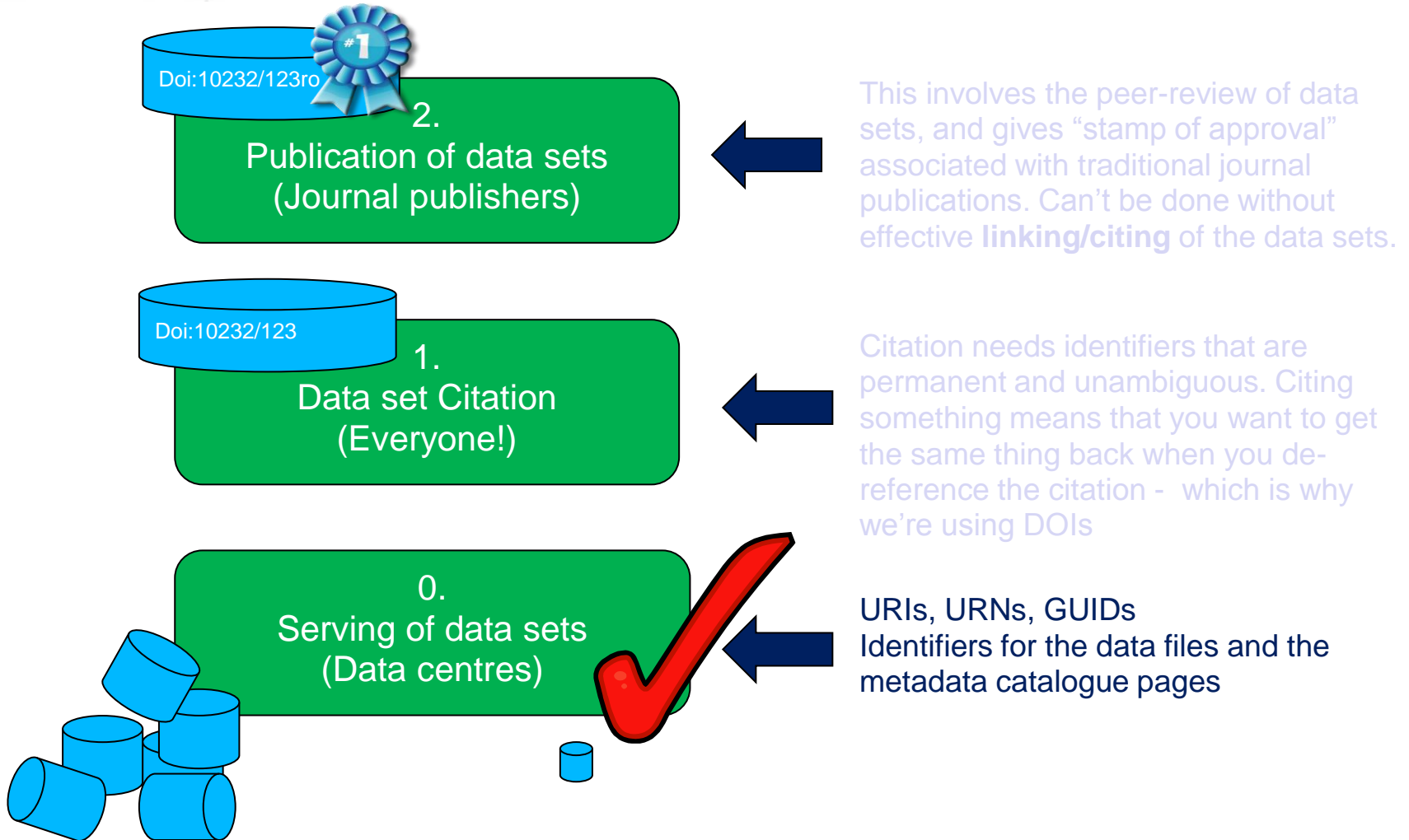
Unless there's a very good reason not to!

Publishing = making something public after some formal process which **adds value** for the consumer: e.g. peer review **and** provides commitment to persistence

Identifiers for data and how we use them



Identifiers for data (2)



- 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!



<http://www.naa.gov.au/records-management/capability-development/keep-the-knowledge/index.aspx>

Out of Cite, Out of Mind: Report of the CODATA Task Group on Data Citation

The report was published by the CODATA Data Science Journal on 13 September 2013



The screenshot shows the J-STAGE website interface. At the top, there is a navigation bar with links for 'About My J-STAGE', 'Sign-in', 'Register', 'Shopping Cart', 'Help', and language options for 'Japanese' and 'English'. Below this is the 'DATA SCIENCE Journal' header, featuring the CODATA logo and the text 'CODATA'. A search bar is located on the right side of the header. The main content area displays the article title 'Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data' by the CODATA-ICSTI Task Group. The article is from 'Data Science Journal', Vol. 12 (2013) p. 1-75. The DOI is <http://dx.doi.org/10.2481/dsj.OSOM13-043>. The article is available as a free PDF (2943K). On the right side, there is a sidebar with 'Article Tools' including 'Add to Favorites', 'Citation Alert', 'Authentication Alert', 'Additional Info Alert', 'Copy the URL', 'Mail to Author', 'Download Meta of Article', 'RIS', 'BibTeX', and 'Contact us'. There is also a 'Share this Article' button at the bottom of the sidebar.

https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article

First Principles for Data Citation

- 1. Status of Data:** Data citations should be accorded the same importance in the scholarly record as the citation of other objects.
- 2. Attribution:** A citation to data should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.
- 3. Persistence:** Citations should refer to objects that persist.
- 4. Access:** Citations should facilitate access to data by humans and by machines.
- 5. Discovery:** Citations should support the discovery of data and their documentation.



First Principles for Data Citation

- 6. Provenance:** Citations should facilitate the establishment of provenance of data.
- 7. Granularity:** Citations should support the finest-grained description necessary to identify the data.
- 8. Verifiability:** Citations should contain information sufficient to identify the data unambiguously.
- 9. Metadata Standards:** Citations should employ existing metadata standards.
- 10. Flexibility:** Citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data across communities..

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.



Viewing GBS 20.7GHz slant x

badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dep_11902119479621181

BADC - Trac METAFOR | Home Google Mail BBC NEWS | News Fr... Sorcha ní gCeallagh... Other bookmarks

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

Search for in All

GBS 20.7GHz slant path radio propagation measurements, Chilbolton site

General Info

Title: GBS 20.7GHz slant path radio propagation measurements, Chilbolton site
Type: Activity
Sub-Type: Deployment
Publication State: Citable
URI: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dep_11902119479621181

Summary

The GBS (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, from August 2003 to August 2006 with signal attenuation sampled once per second.

Please cite this dataset as:

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S. A. Callaghan, J. Waight, C. J. Walden, J. Agnew and S. Ventouras]. GBS 20.7GHz slant path radio propagation measurements, Sparsholt site, [Internet]. British Atmospheric Data Centre, 2003-2005, 1st April 2011, doi:10.5285/639a3714-bc74-46a6-9026-64931f355e07

This dataset is cited in:

S. A. Callaghan, J. Waight, J.L.Agnew, C. J. Walden, C.L.Wrench, S. Ventouras "The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK", Geoscience Data Journal, 17 March 2013, DOI: 10.1002/gdj3.2

Author

Name email

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S. A. Callaghan, J. Waight, C. J. Walden, J. Agnew and S. Ventouras]

Online References

Relation	Title
Apply for access	Apply for to GBS data from Chilbolton
Download	Data directory for GBS data from Chilbolton
Documentation	DOI for dataset 10.5285/639a3714-bc74-46a6-9026-64931f355e07
Documentation	Data article in Geoscience Data Journal doi:10.1002/gdj3.2

Associated Data

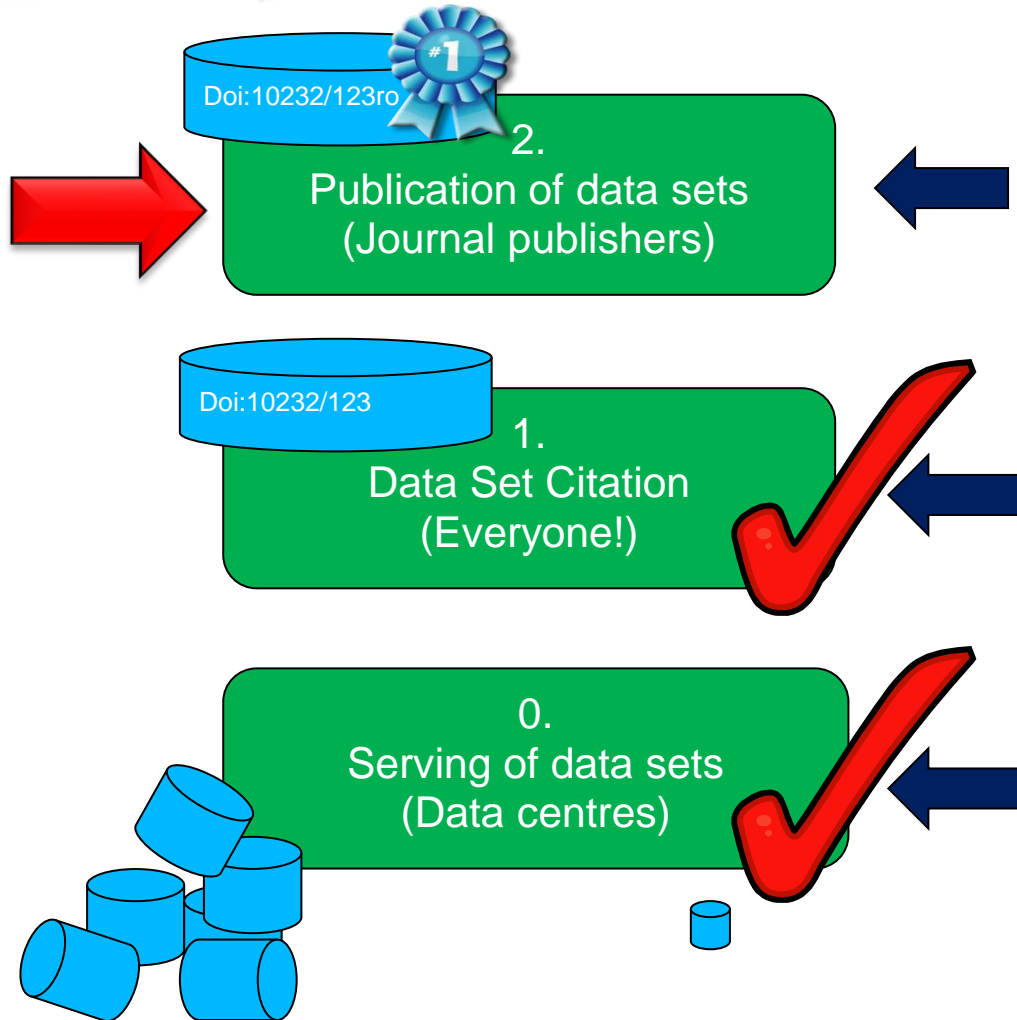
Type	Title
Data Production Tool	Chilbolton: GBS receiver
Activity	Chilbolton Facility for Atmospheric and Radio Research (CFARR)
Observation Station	Chilbolton Facility for Atmospheric and Radio Research (CFARR), UK

Dataset catalogue page (and DOI landing page)

Dataset citation

Clickable link to Dataset in the archive

Foundations and links are in place – now what?



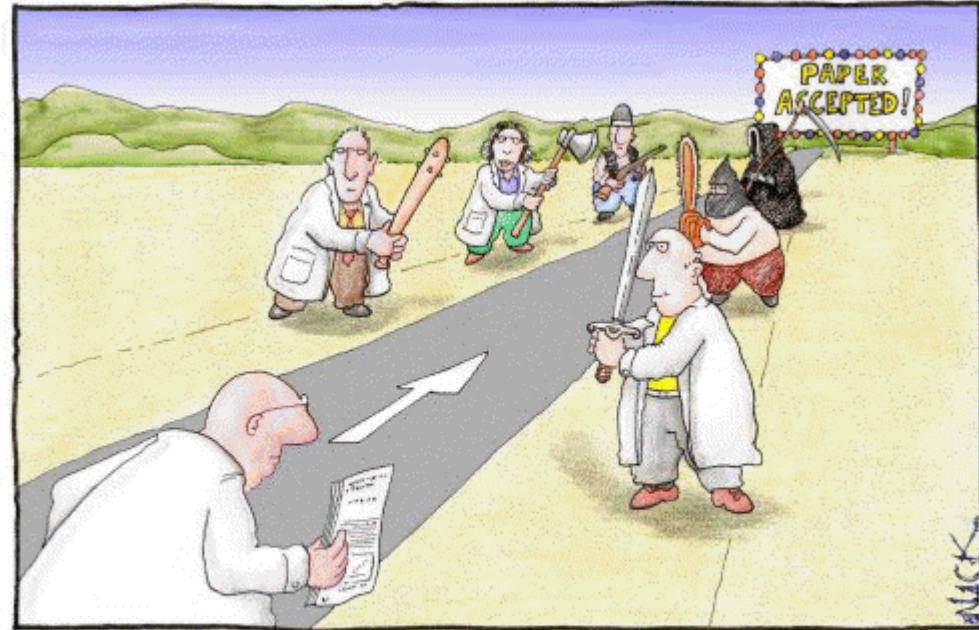
The **scientific quality** of a dataset has to be evaluated by **peer-review** by scientists with domain knowledge. This peer-review process has already been set up by academic publishers, so it makes sense to collaborate with them for peer-review publishing of data.

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.

Peer review, data and data journals

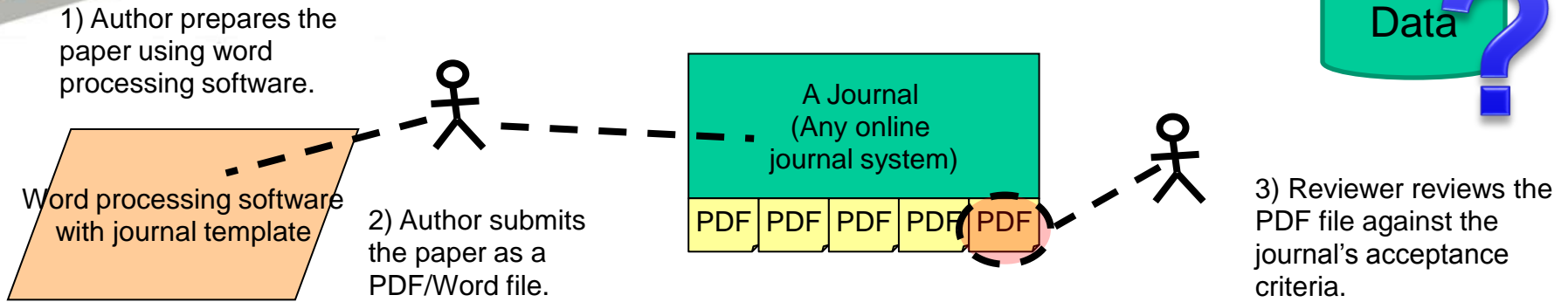
- **Peer-review** of a scientific publication is generally only applied to **analysis, interpretation and conclusions**, and **not the underlying data**.
- But if the conclusions are valid, the **data must be of good quality**.
- We need **quality assurance** of the data underlying research publications – either through peer-review or data repository checking.
- Researchers need **credit** for creating, managing and opening their data.
- Data journals provide that credit in an environment where academic status is solely based on publication record.



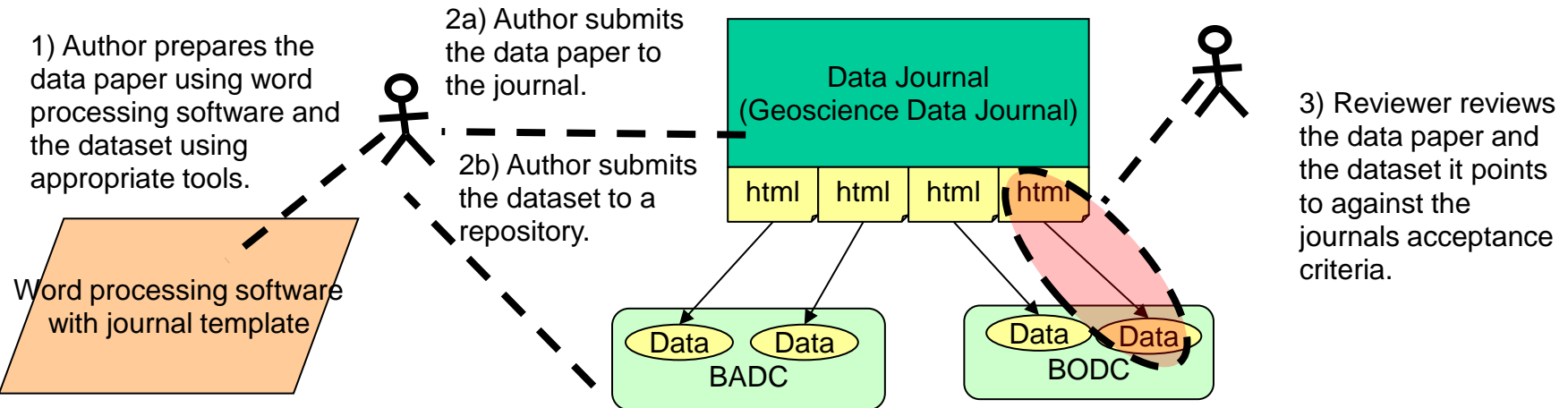
Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

<http://libguides.luc.edu/content.php?pid=5464&sid=164619>

The traditional online journal model



Overlay journal model for publishing data



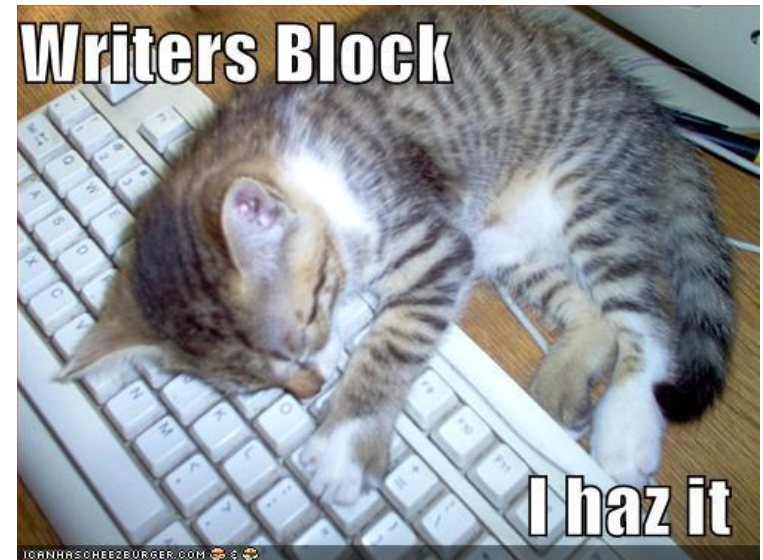
What is a data article?

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

- the **when, how and why** data was collected and what the data-product is.

Many data journals already exist – see a list (in no particular order) at:

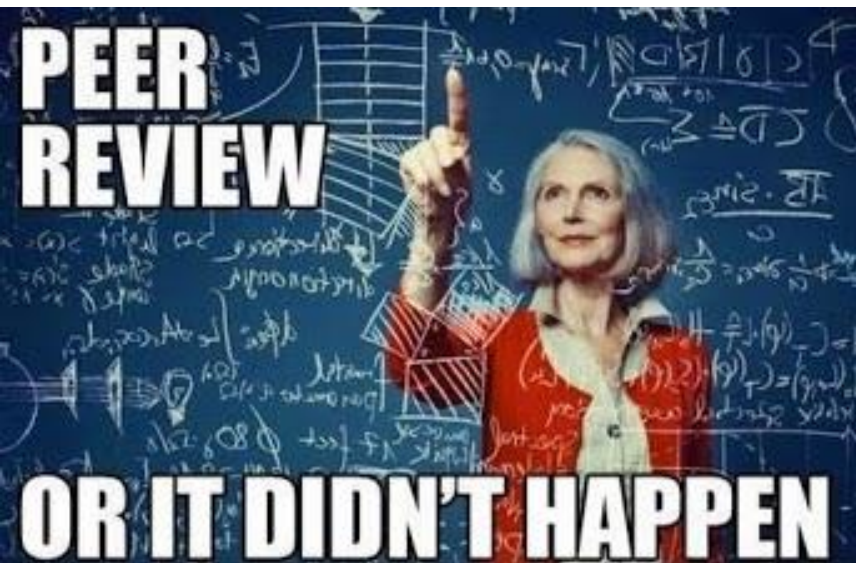
<http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList>



Why bother publishing the dataset in a data journal? Why not just publish a normal journal paper citing the data?

Data Journals:

- Peer-review the data
- Publish negative results
- Make it quicker to publish the data as they don't require analysis or novelty – the dataset is published “as-is”
- Provide attribution and credit for the data collectors who might not be involved with the analysis
- Make it easier to find datasets, understand them and be sure of their quality and provenance.



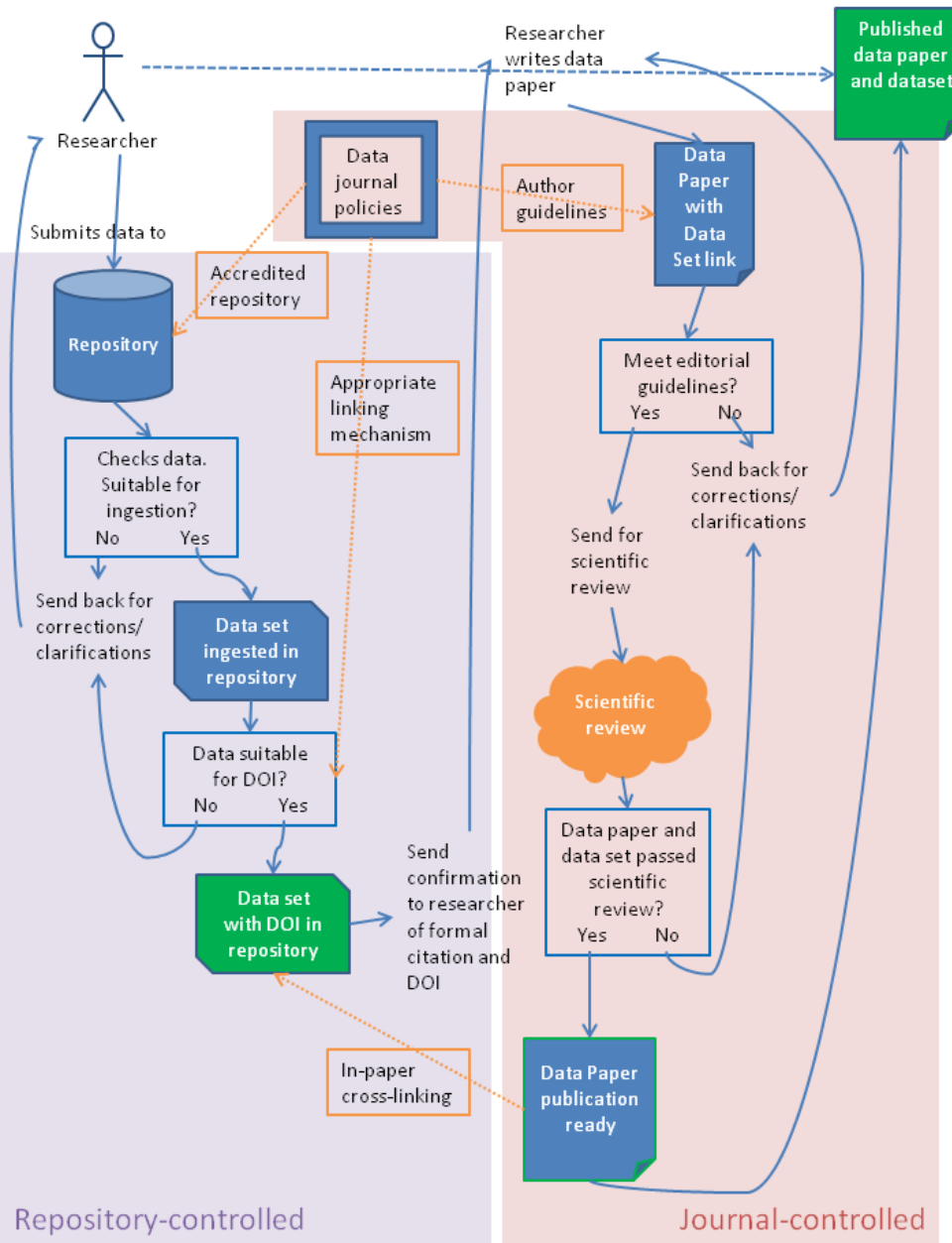
PREPARDE: Peer REVIEW for Publication & Accreditation of Research Data in the Earth sciences

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
<http://bit.ly/ZhYHZI>
 3. Scientific peer-review of data
<http://bit.ly/DataPRforComment>
- Division of area of responsibilities between
 - *repository controlled* processes
 - *journal controlled* processes

<http://proj.badc.rl.ac.uk/preparde/wiki>



The GBS dataset: measure x

onlinelibrary.wiley.com/doi/10.1002/gdj3.2/full

BADC - Trac | METAFOR | Home | Google Mail | BBC NEWS | News Fr... | Sorcha ní gCeallagh... | Add to Wish List | Other bookmarks

WILEY ONLINE LIBRARY

PUBLICATIONS | BROWSE BY SUBJECT | RESOURCES | ABOUT US | The Chadwick & RAL Libraries

Home > Earth Sciences > General & Introductory Earth Sciences > Geoscience Data Journal > Early View > Abstract

JOURNAL TOOLS

- Get New Content Alerts
- Get RSS feed
- Save to My Profile
- Recommend to Your Librarian

JOURNAL MENU

Journal Home

FIND ISSUES

Early View

FIND ARTICLES

Early View

FOR CONTRIBUTORS

Author Guidelines

Submit an Article

ABOUT THIS JOURNAL

Society Information

Overview

Editorial Board

Permissions

SPECIAL FEATURES

Data Center FAQs

L.F. Richardson Award Prize Winners

Open Access License and Copyright

Author FAQs

Article Publication Charges

Wiley Open Access

Institutional and Funder Payments

Guidelines for Reviewers

Guidelines for Repositories

RMetS Geoscience Data Journal

Open Access

Data Paper

The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK

S. A. Callaghan¹, J. Waight, J. L. Agnew, C. J. Walden, C. L. Wrench, S. Ventouras

Issue

Article first published online: 17 MAR 2013
DOI: 10.1002/gdj3.2

Copyright © 2013 The Authors. Published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes

Geoscience Data Journal
Early View (Online Version of Record published before inclusion in an issue)

SEARCH

In this issue

Advanced > Saved Searches >

ARTICLE TOOLS

- Get PDF (359K)
- Save to My Profile
- E-mail Link to this Article
- Export Citation for this Article
- Get Citation Alerts
- Request Permissions

Share |

Additional Information (Show All)

How to Cite | Author Information | Publication History | Funding Information

The research presented in this paper was funded by the UK's Ofcom as part of the Spectrum Efficiency Scheme and the support of Ofcom in providing the funding for the GBS experiment is greatly appreciated.

Abstract | Article | References | Cited By

Get PDF (359K)

Keywords:
site diversity; radio propagation; fade mitigation techniques

Abstract

The GBS (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 20.7 GHz 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.

Dataset

The GBS (Global Broadcast Service) dataset comes as 3 separate data streams:

- Identifier: [doi:10.5285/639A3714-BC74-46A6-9026-64931F355E07](https://doi.org/10.5285/639A3714-BC74-46A6-9026-64931F355E07)
Creator: Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].
Title: GBS 20.7 GHz slant path radio propagation measurements, Chilbolton site
publisher: NERC British Atmospheric Data Centre
Publication year: 2009
Resource type: Metadata document
Version: 1.0
- Identifier: [doi:10.5285/db8d8981-1a51-4d6e-81c0-cced9b921390](https://doi.org/10.5285/db8d8981-1a51-4d6e-81c0-cced9b921390)
Creator: Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan, S. A., J. Waight, C. J. Walden, J. Agnew and S. Ventouras].

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

Viewing GBS 20.7GHz slant path radio propagation measurements, Chilbolton site

badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dep_11902119479621181

BADC - Trac METAFOR | Home Google Mail BBC NEWS | News Fr... Sorcha ní gCeallagh... Other bookmarks

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

Search for in All

GBS 20.7GHz slant path radio propagation measurements, Chilbolton site

General Info

Title: GBS 20.7GHz slant path radio propagation measurements, Chilbolton site
Type: Activity
Sub-Type: Deployment
Publication State: Citable
URI: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dep_11902119479621181

Summary

The GBS (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, from August 2003 to August 2006 with signal attenuation sampled once per second.

Please cite this dataset as:
 Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S. A. Callaghan, J. Waight, C. J. Walden, J. Agnew and S. Ventouras]. GBS 20.7GHz slant path radio propagation measurements, Sparsholt site, [Internet]. British Atmospheric Data Centre, 2003-2005. 1st April 2011. doi:10.5285/628-2714-1-71-46x6-0026-64821f355e07

This dataset is cited in:
 S. A. Callaghan, J. Waight, J.L.Agnew, C. J. Walden, C.L.Wrench, S. Ventouras "The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK", Geoscience Data Journal, 17 March 2013, DOI: 10.1002/gdj3.2

Author

Name email
 Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [S. A. Callaghan, J. Waight, C. J. Walden, J. Agnew and S. Ventouras]

Online References

Relation	Title
Apply for access	Apply for to GBS data from Chilbolton
Download	Data directory for GBS data from Chilbolton
Documentation	DOI for dataset:10.5285/628-2714-1-71-46x6-0026-64821f355e07
Documentation	Data article in Geoscience Data Journal doi:10.1002/gdj3.2

Associated Data

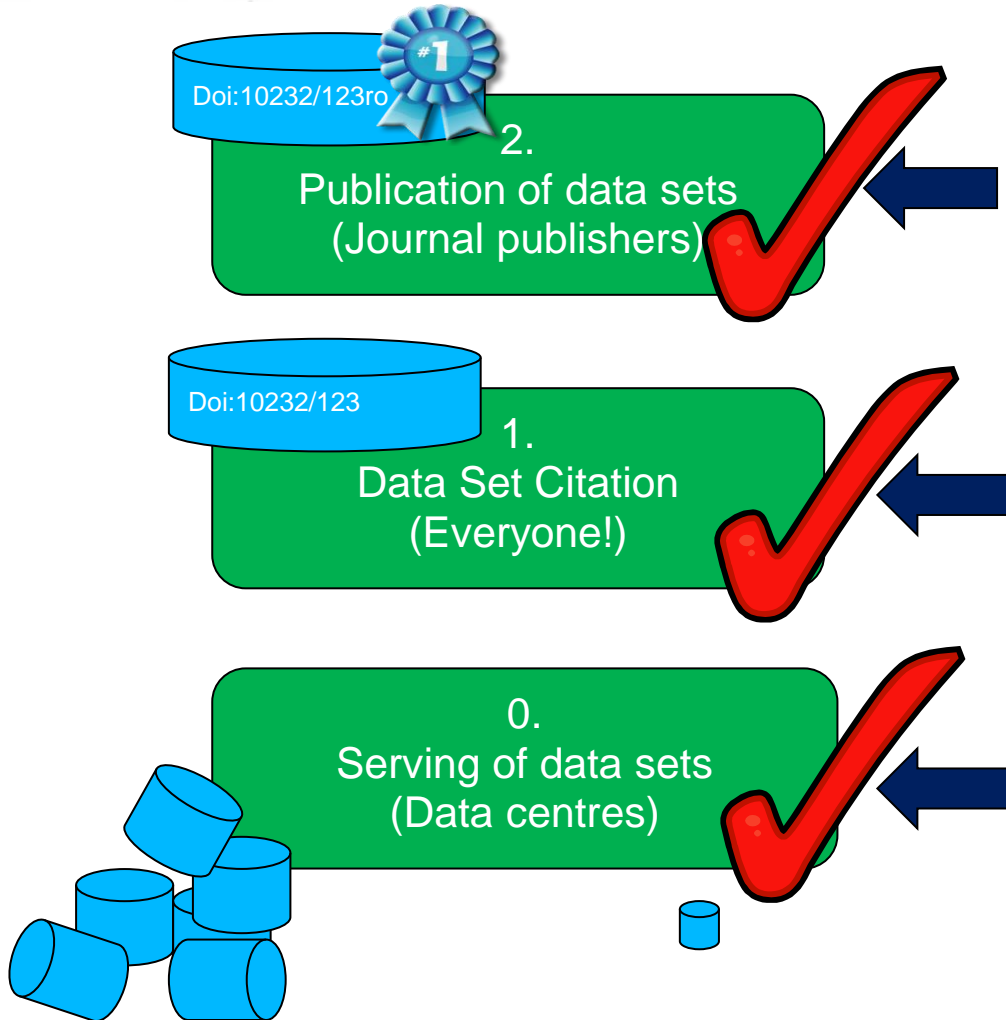
Type	Title
Data Production Tool	Chilbolton: GBS receiver
Activity	Chilbolton Facility for Atmospheric and Radio Research (CFARR)
Observation Station	Chilbolton Facility for Atmospheric and Radio Research (CFARR), UK

Dataset catalogue page (and DOI landing page) – again!

Reference to Data Article

Clickable link to Data Article

What we've done and how we've done it



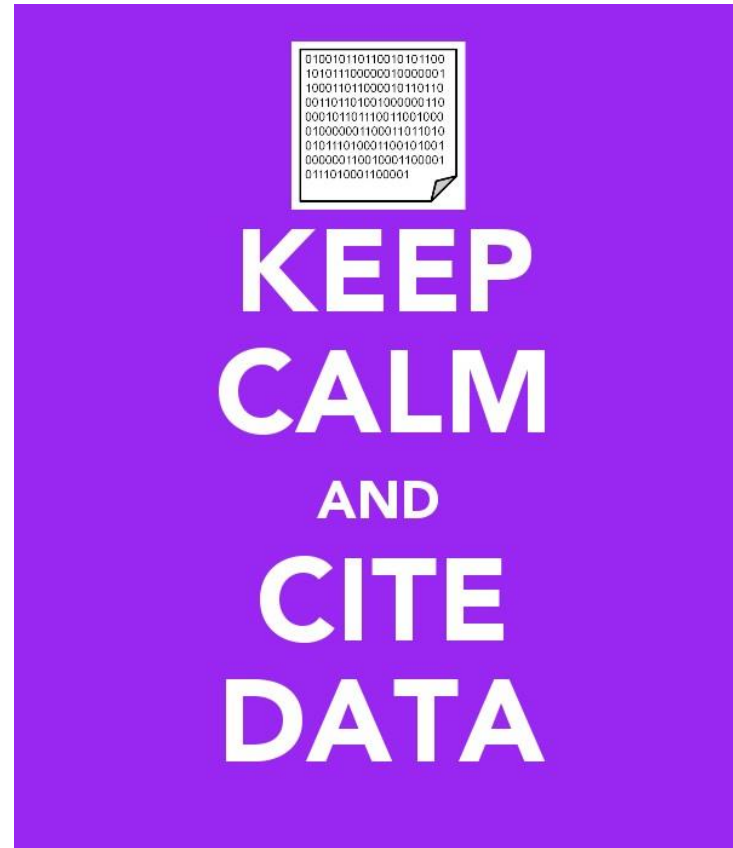
Data paper has been published in a data journal, linked via DOI to underlying dataset. Formal citations of datasets (also using DOIs) done in standard academic articles.

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 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 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're progressing well with data publication through our partnership with Wiley-Blackwell, and discussions with Elsevier. 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-omatic.co.uk/default.aspx#createposter>

Cost Action: Publishing Academic and Research Data (PARD)

COST is a mechanism in the EU to fund networking activities on topics in science and technology – meetings, workshops, short term scientific missions...bringing people together

>50 people interested

13 countries including: United Kingdom, Austria, Germany, Poland, Hungary, France, Italy, Netherlands, Bulgaria, Norway, Slovenia, Sweden, USA

For more information – or to join!

Sarah Callaghan
[sarah.callaghan@stfc.ac.uk]
@sorcha_ni

A **Pard** is an animal from Medieval bestiaries. They were felines with spotted coats, and were extremely fast.



<http://en.wikipedia.org/wiki/File:AberdeenBestiaryFolio008vLeopardDetail.jpg>

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>

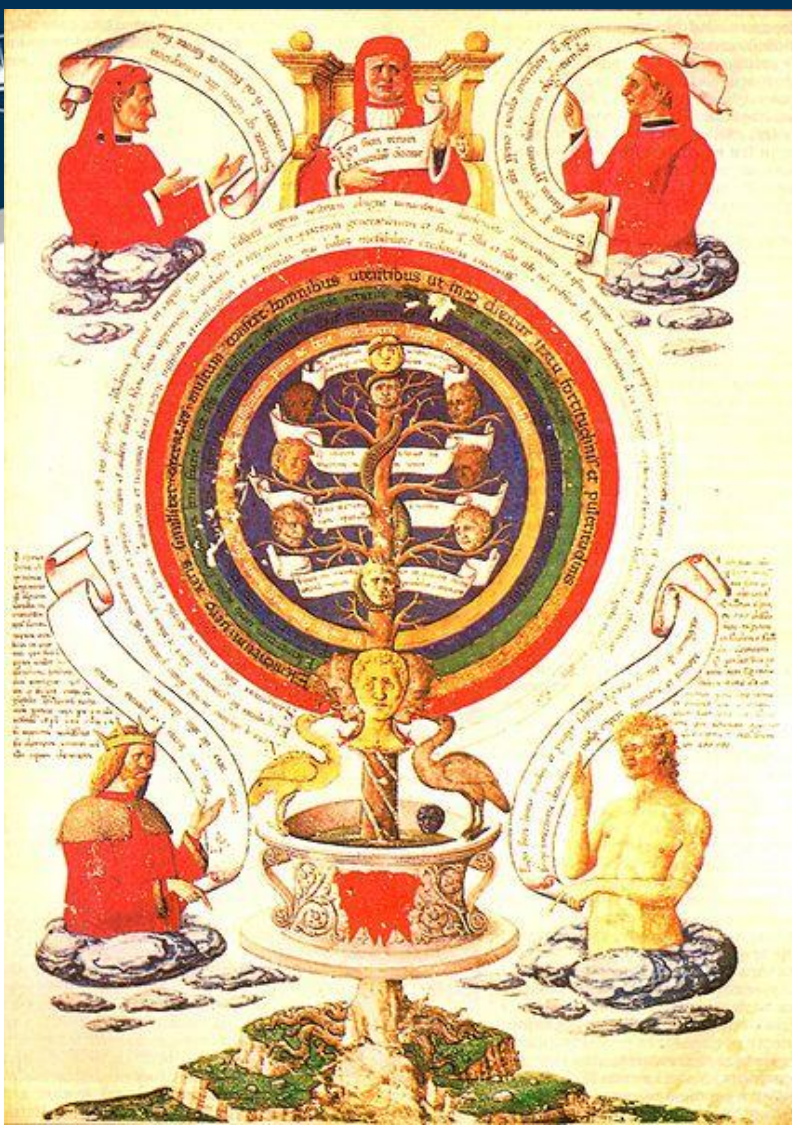
Guidelines on peer review for data:

<http://bit.ly/DataPRforComment>

Guidelines for repository accreditation for data publication: <http://bit.ly/ZhYHZI>

Feedback to:

<https://www.jiscmail.ac.uk/DATA-PUBLICATION>



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

http://en.wikipedia.org/wiki/File:Raimundus_Lullus_alchemic_page.jpg