Science & Technology Facilities Council

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

Sarah Callaghan*
sarah.callaghan@stfc.ac.uk
@sorcha.ni

*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

British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

NGDC
National Geoscience Data Centre

BODC

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NEODC

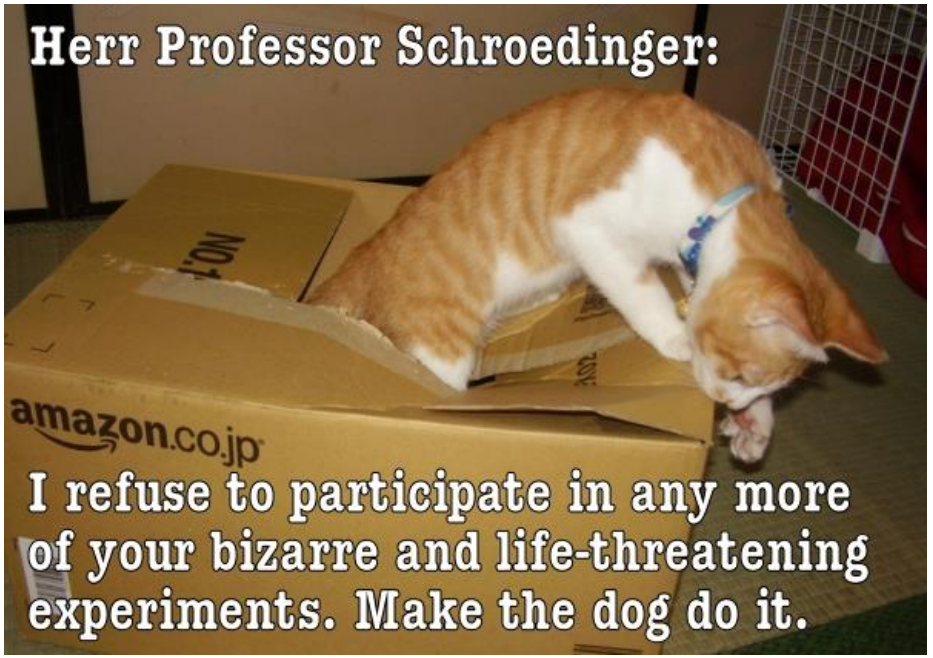
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

Data, Reproducibility and Science

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

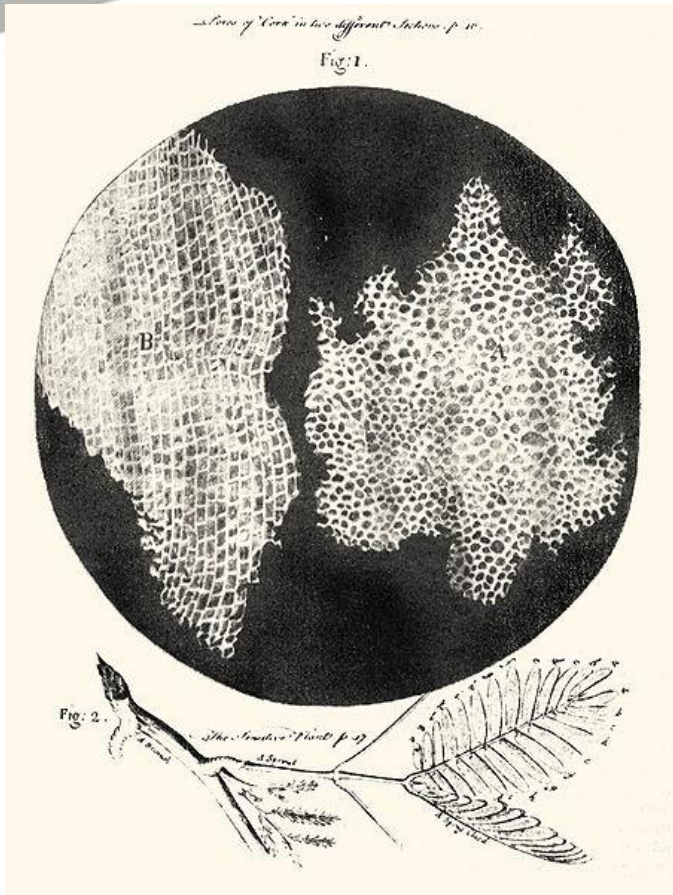
Observational data is not reproducible (unless you have a time machine!)

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		
	n f	.....	237 31	3		
	a f	.....	335 23	3		
	a g	*g* =(12.13)	215 17	3	115·5	4
	a k	*k* =(12.13)	193 29	3		
	g k	.....	87 5	3		
May 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	355·2	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

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

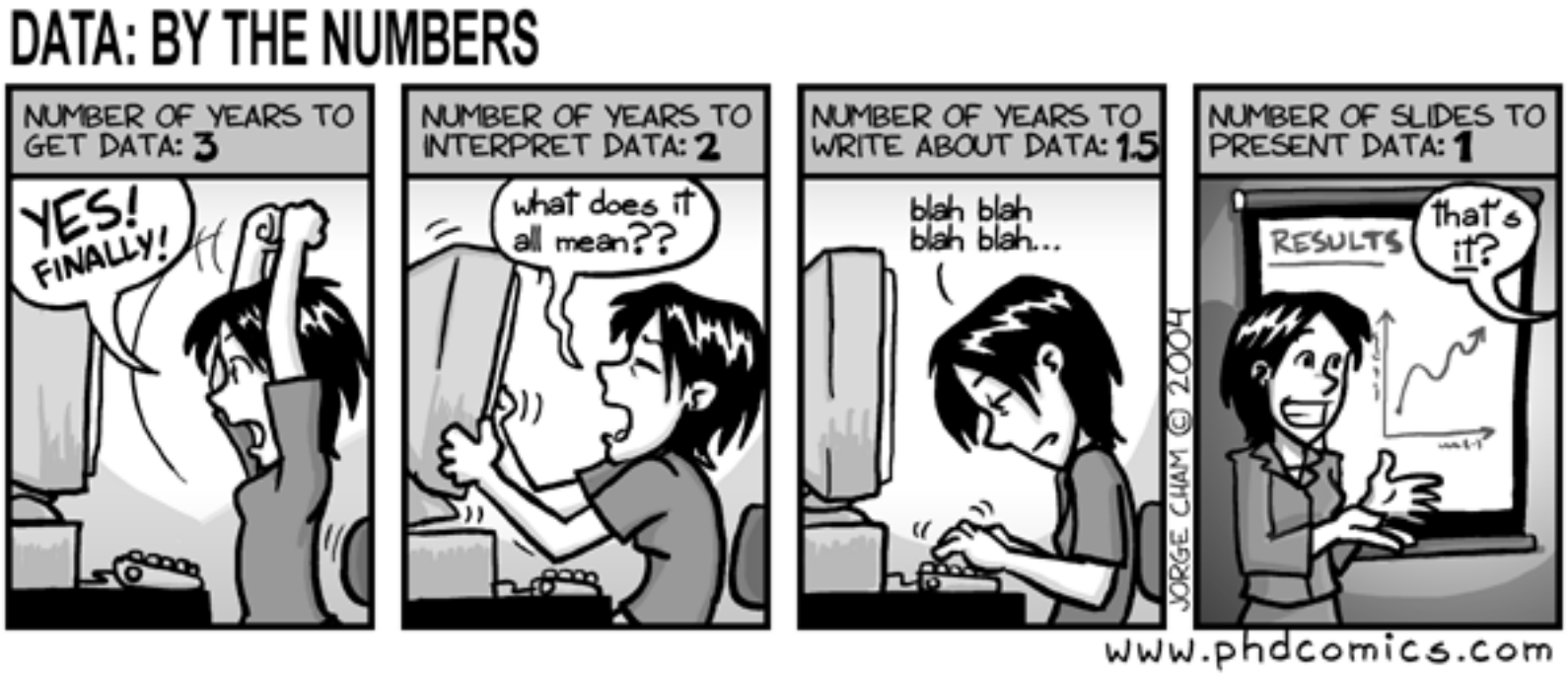
- Pressure from (UK) government to make data from publicly funded research available for free.
  - Scientists want attribution and credit for their work
  - Public want to know what the scientists are doing
  - Good for the economy if new industries can be built on scientific data/research

- 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

Creating a dataset is hard work!



"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.

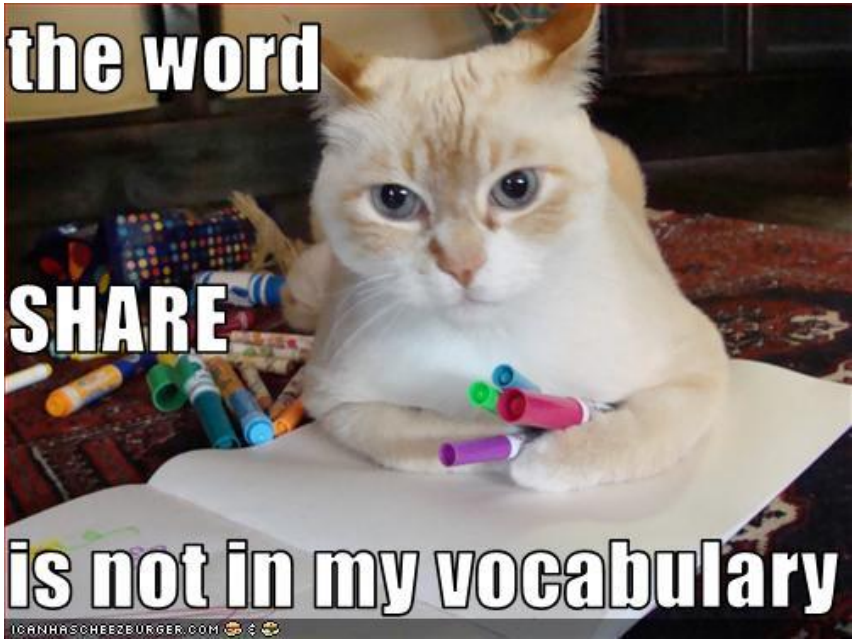
Knowledge is power!

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.



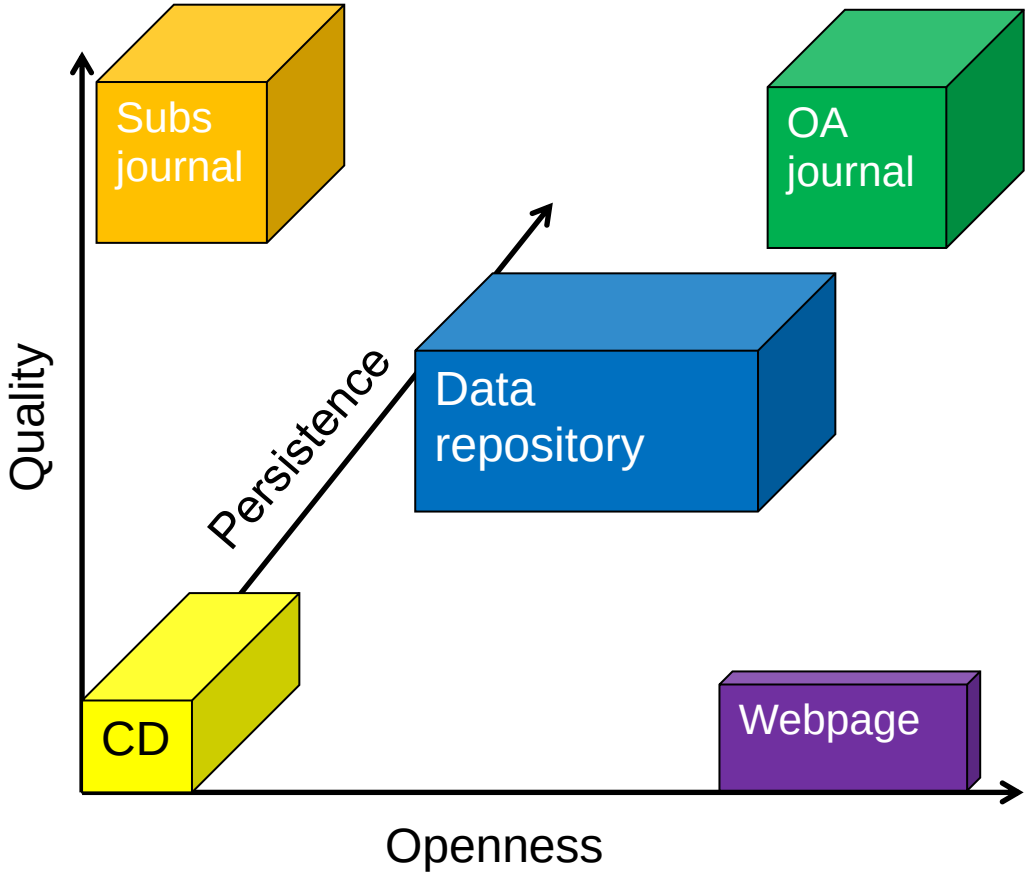
But if we publish the data, precedence is established and credit is given!

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



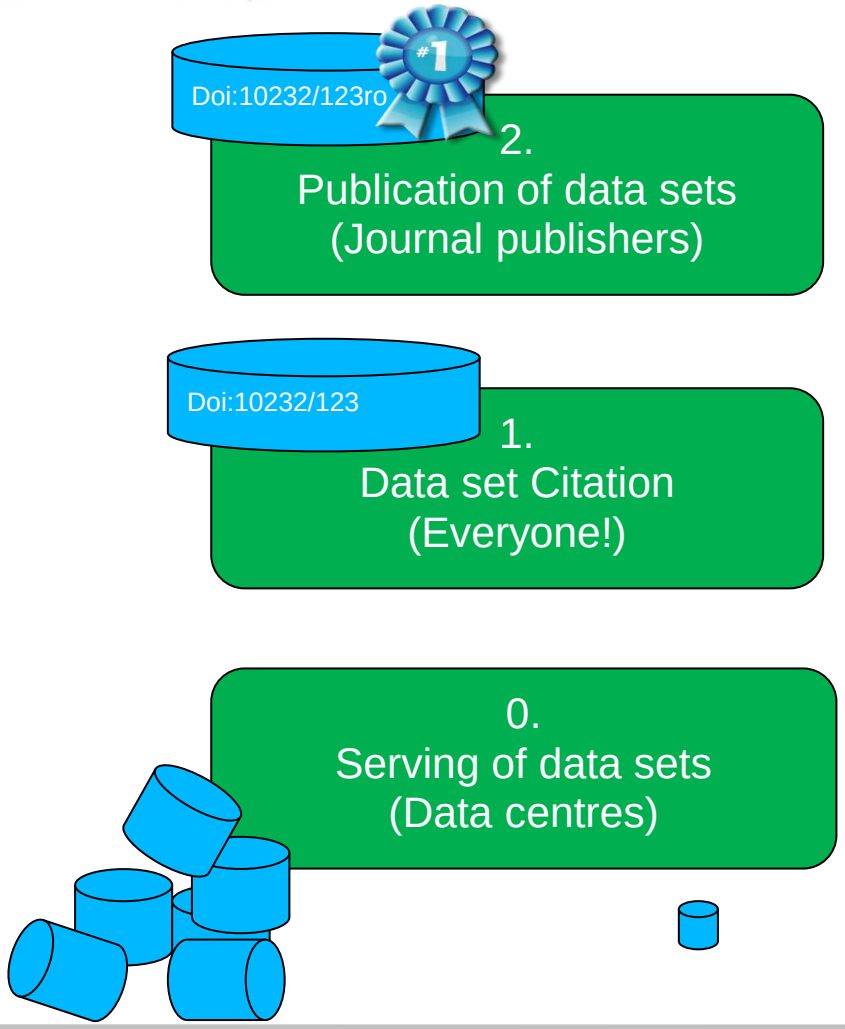
We want to encourage researchers to make their data:

- 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

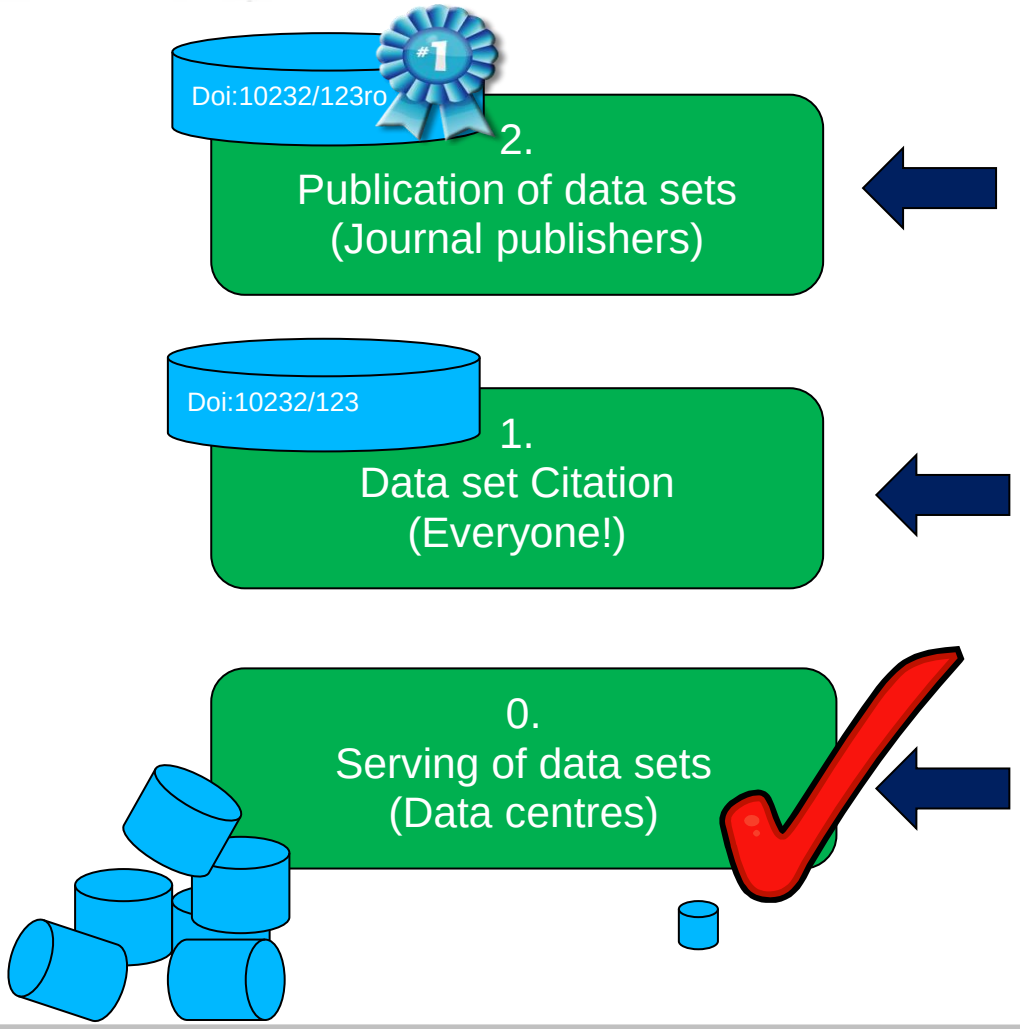


This involves the peer-review of data sets, and gives "stamp of approval" associated with traditional journal publications. Can't be done without effective **linking/citing** of the data sets.

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!

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

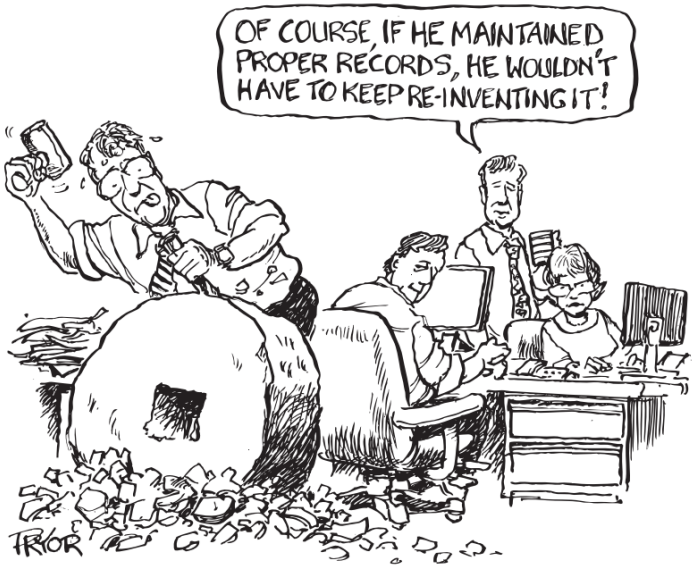
URIs, URNs, GUIDs
Identifiers for the data files and the metadata catalogue pages

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!



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

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.

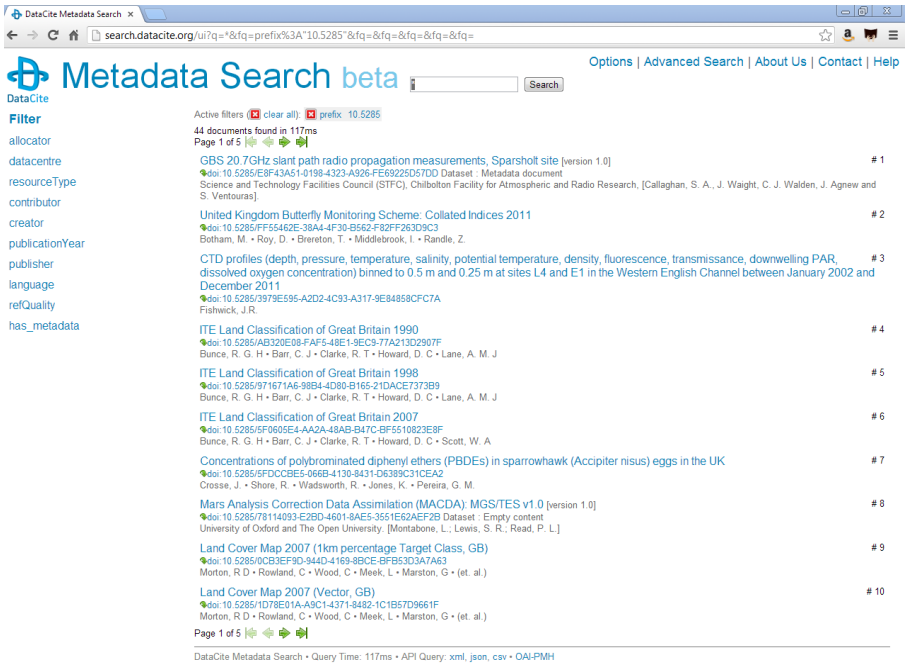
10. Flexibility: Citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data across communities..

Science & Technology Facilities Council

How we (NERC) 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)
- Academic journal publishers are starting to require datasets be cited in a stable way, i.e. using DOIs.
- We have a good working relationship with the British Library and DataCite



What sort of data can we/will we assign a DOI to?

Dataset has to be:

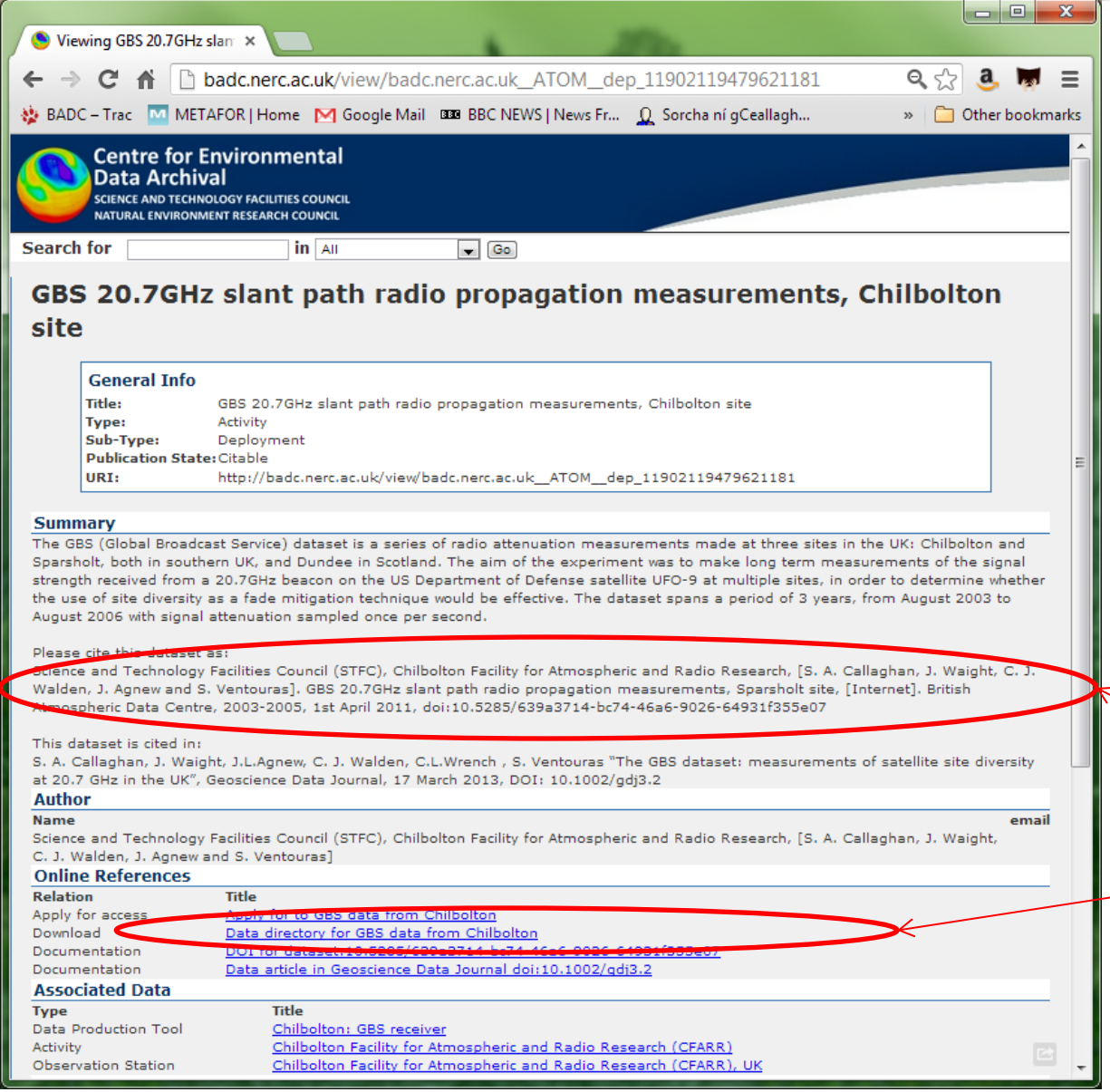
- Stable (i.e. not going to be modified)
- Complete (i.e. not going to be updated)
- Permanent – by assigning a DOI we're committing to make the dataset available for posterity
- Good quality – by assigning a DOI we're giving it our data centre stamp of approval, saying that it's complete and all the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



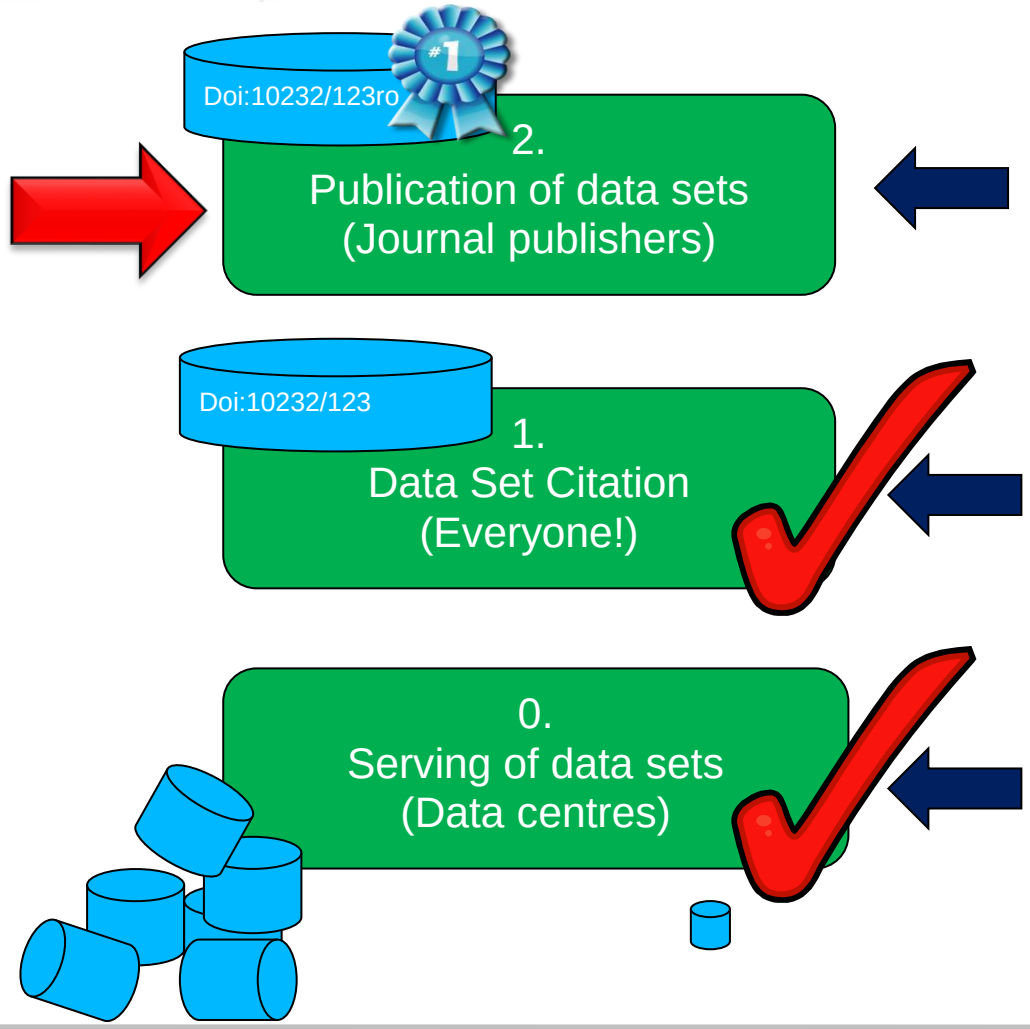
Dataset catalogue page (and DOI landing page)

Dataset citation

Clickable link to Dataset in the archive

Science & Technology Facilities Council

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 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.

Peer review, data and data journals

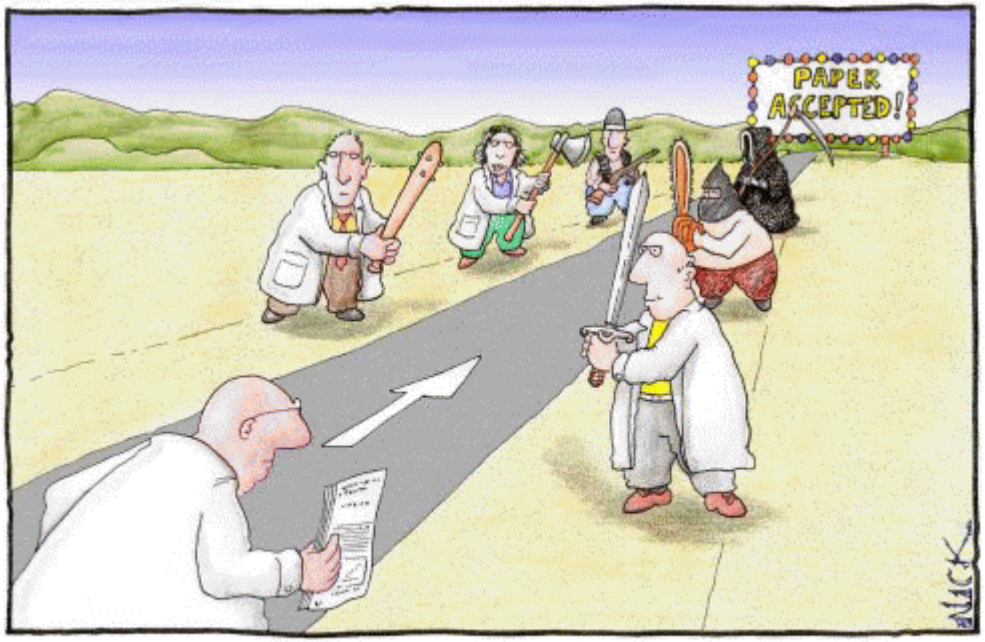
- Peer-review of a scientific publication is generally only applied to analysis, interpretation and conclusions, and not the underlying data.

- 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.

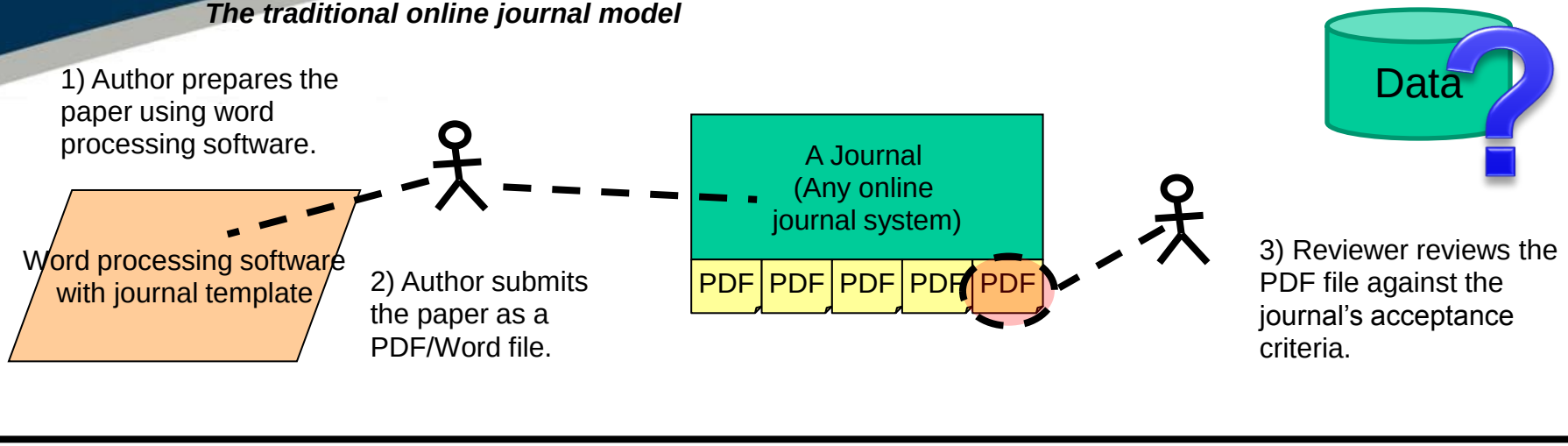
- Data journals provide that credit in an environment where academic status is solely based on publication record.



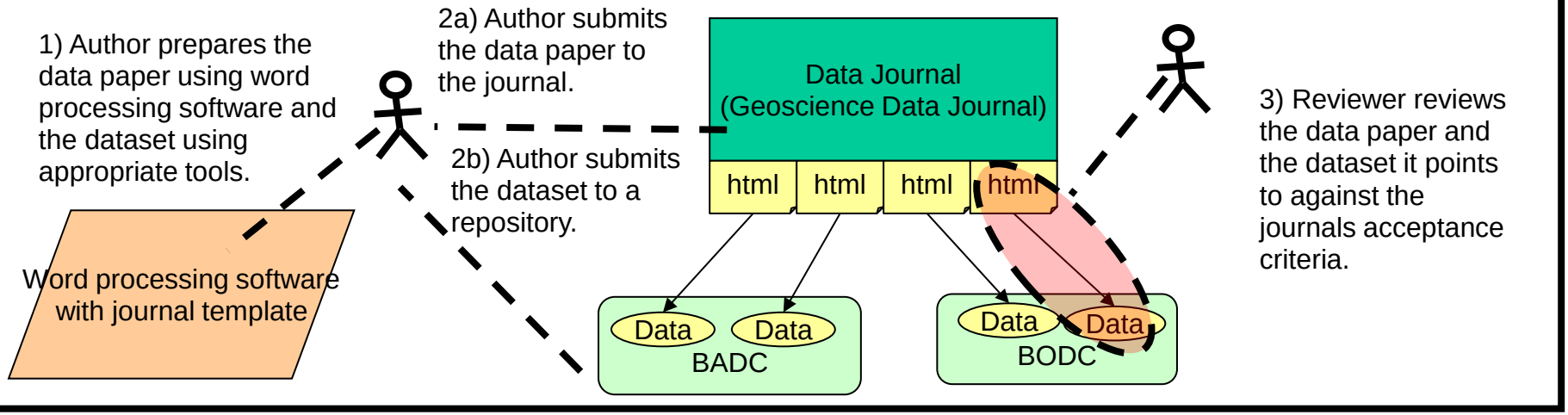
Most scientists regarded the new streamlined peer-review process as 'quite an improvement.'

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.

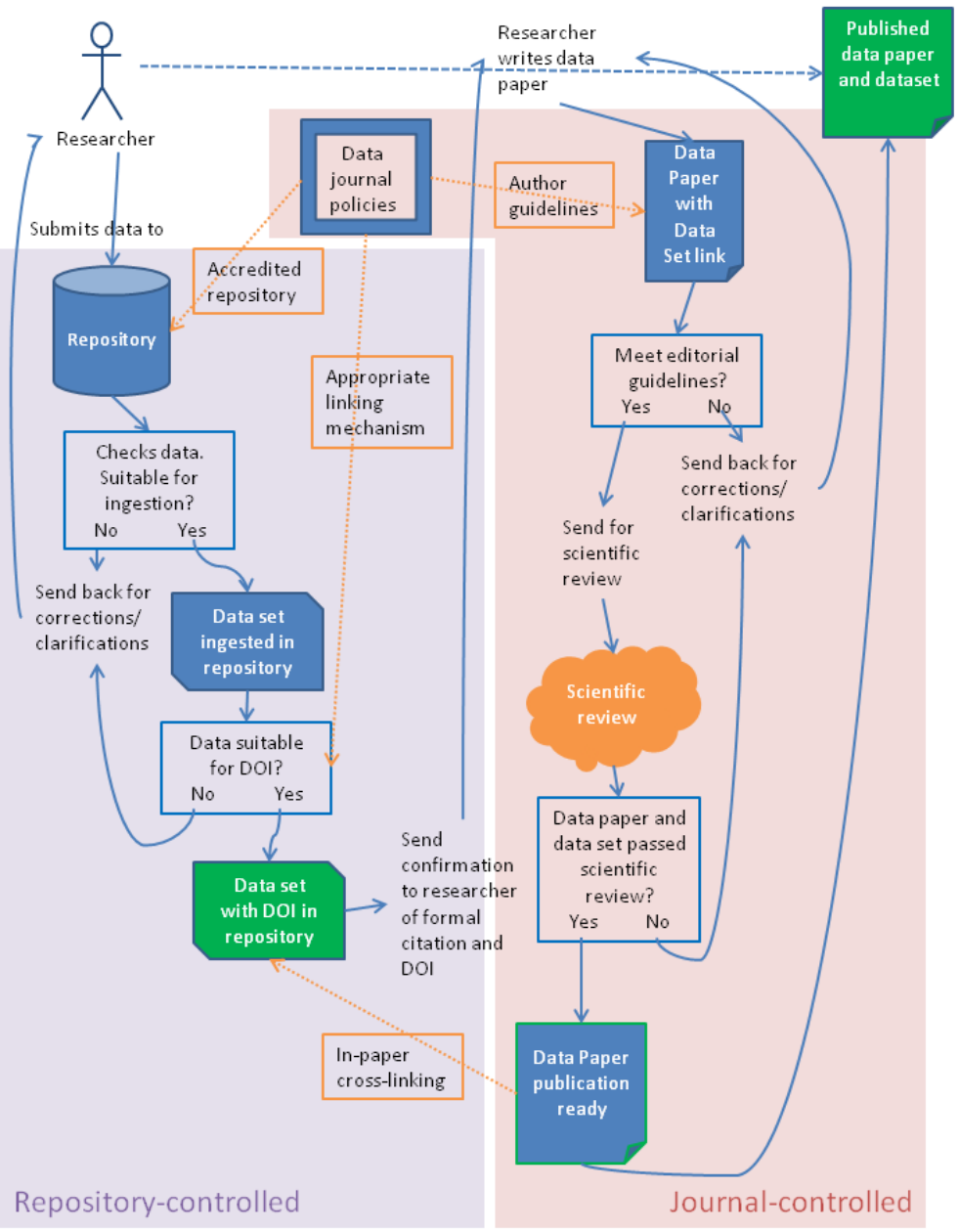
- the when, how and why data was collected and what the data-product is.

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

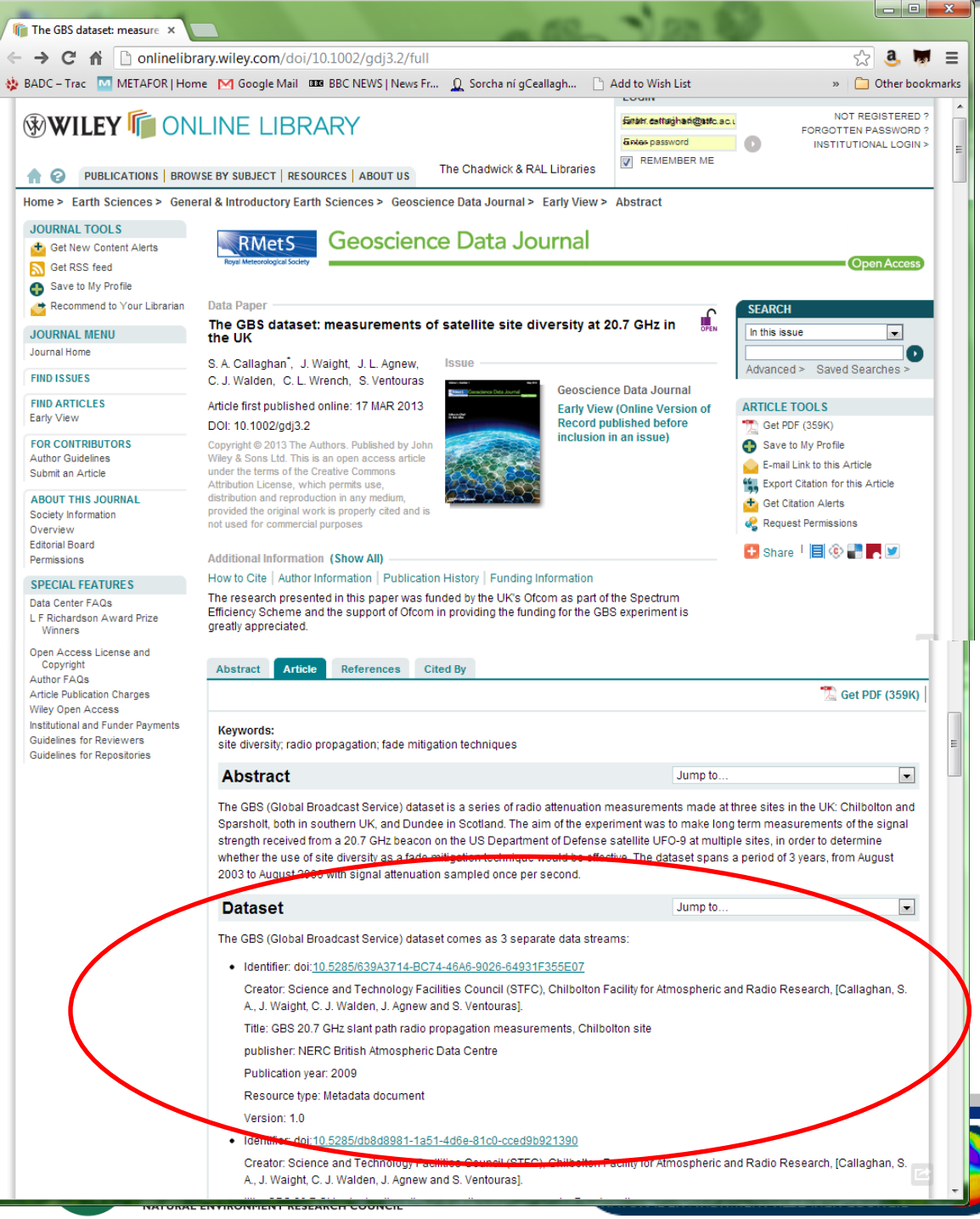
Example steps/workflow required for a researcher to publish a data paper

3 main areas of interest (in orange)

1. Workflows and cross-linking between journal and repository
2. Repository accreditation http://bit.ly/ZhYHZl
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

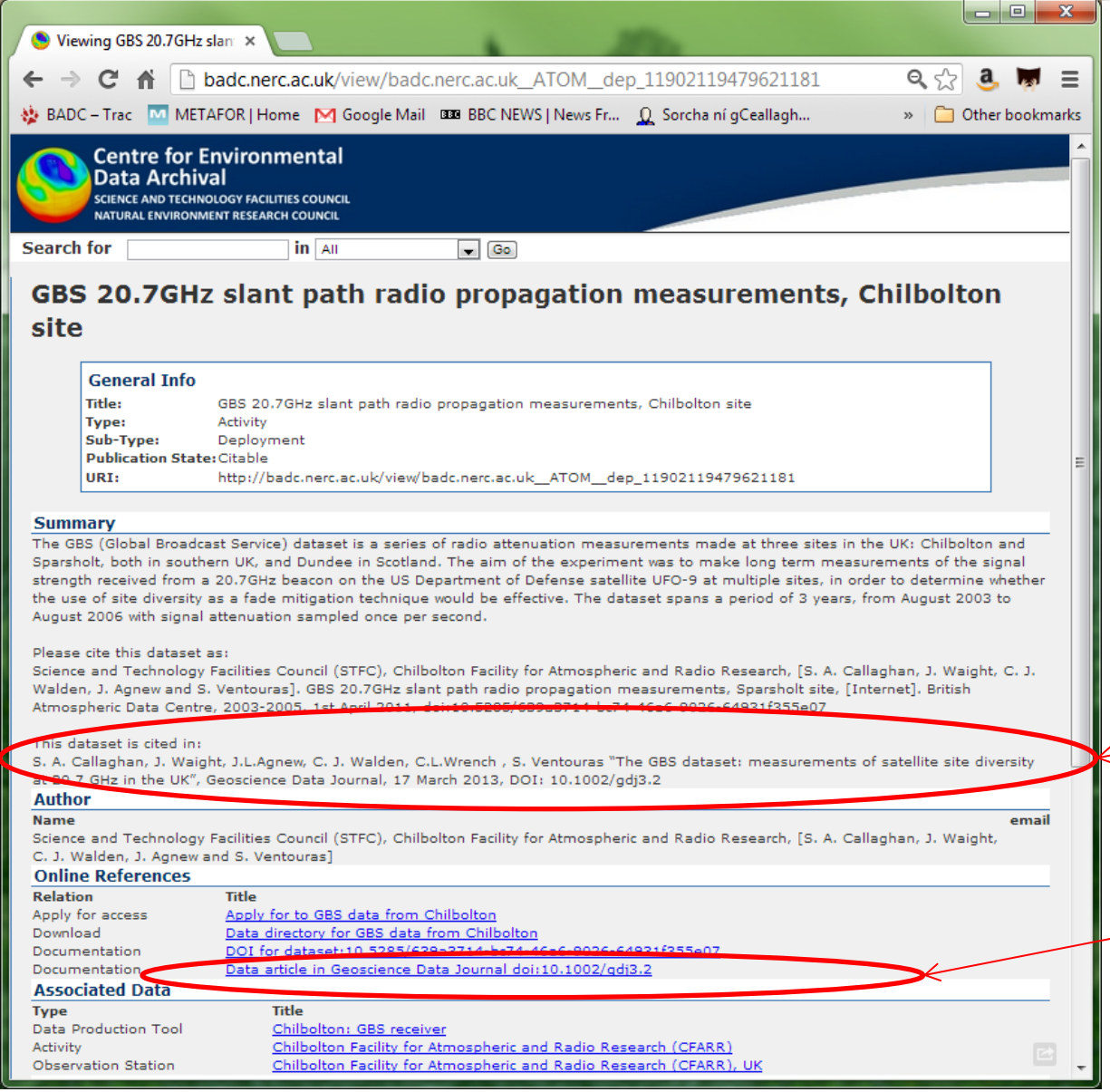


Live Data Paper in *Geoscience Data Journal*!

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

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

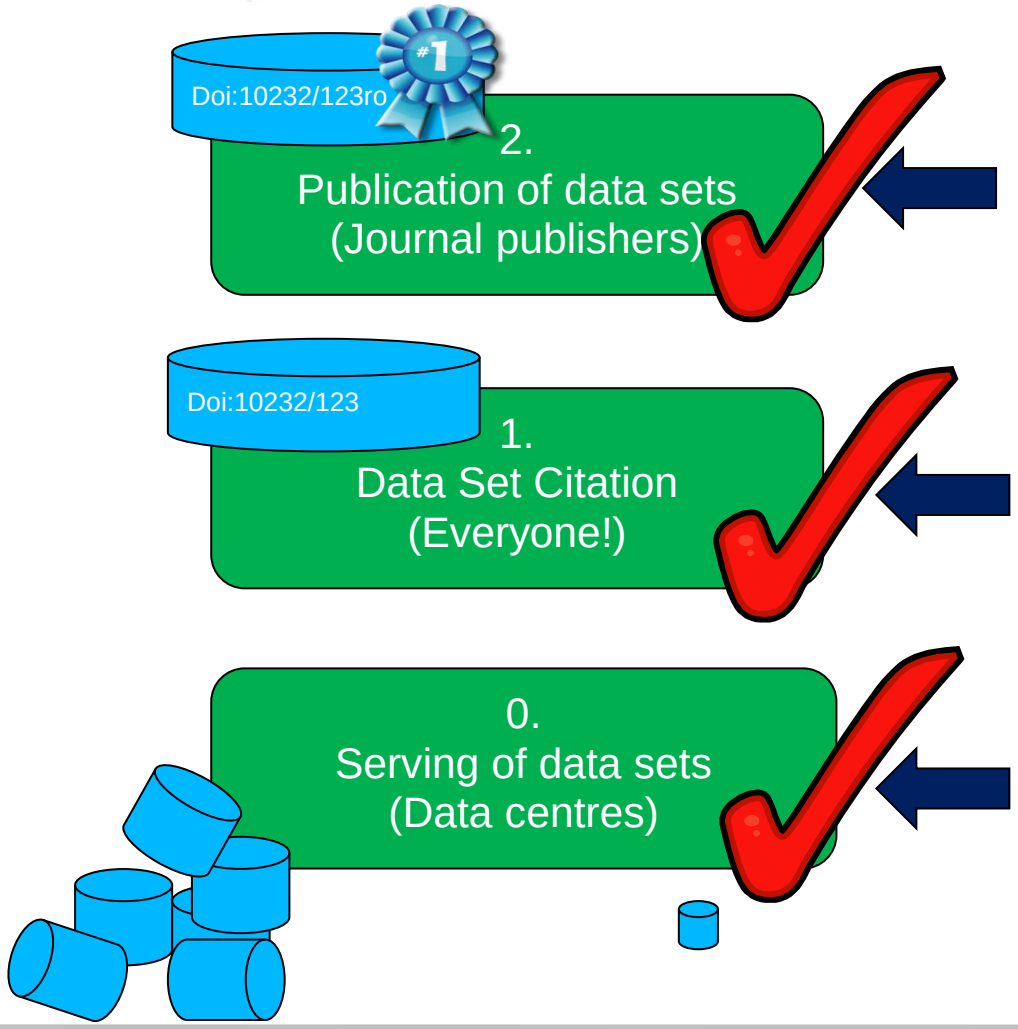


Dataset catalogue page (and DOI landing page) – again!

Reference to Data Article

Clickable link to Data Article

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



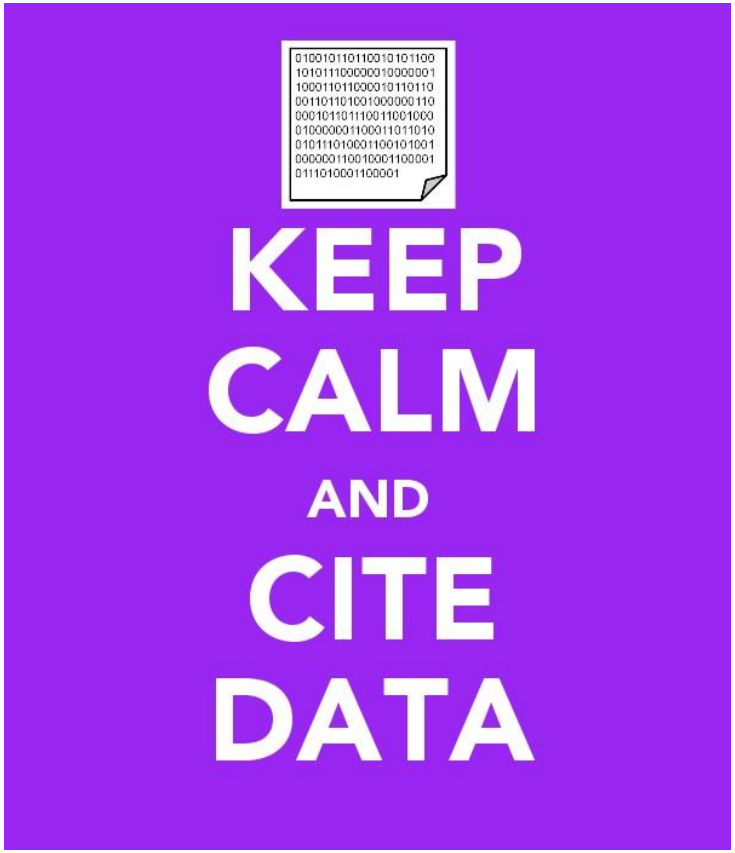
Data paper has been published in a data journal, linked via DOI to underlying dataset. Formal citations of datasets (also using DOIs) done in standard academic articles.

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-o-matic.co.uk/default.aspx#createposter

Science & Technology Facilities Council

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

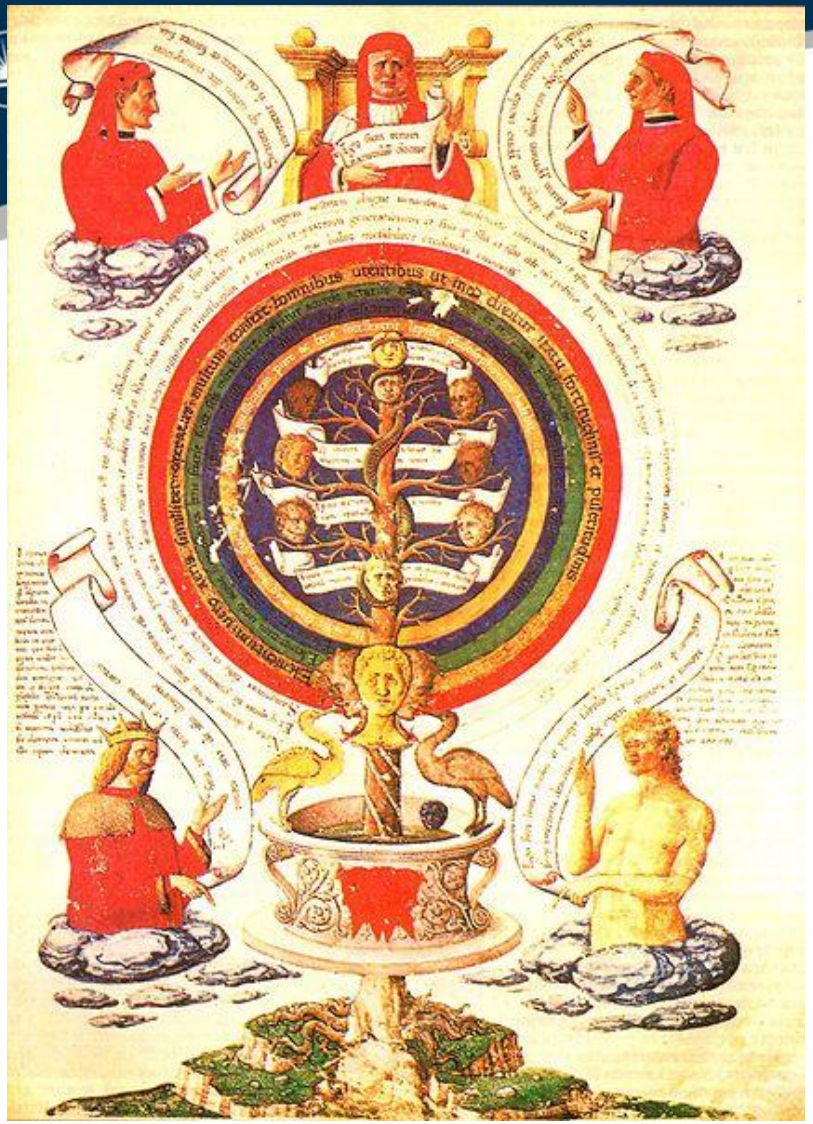
13 countries including: United Kingdom, Austria, Germany, Poland, Hungary, France, Italy, Netherlands, Bulgaria, Norway, Slovenia, Sweden, USA

For more information – or to join!

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

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

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



Page from alchemic treatise of Ramon Llull (Beginning of the 16th century)
http://en.wikipedia.org/wiki/File:Raimundus_Lullus_alchemic_page.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/ZhYHZl
Feedback to: https://www.jiscmail.ac.uk/DATA-PUBLICATION