Science & Technology Facilities Council

Data Citation, Principles and Practice

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

DataCite Annual Conference, 2014

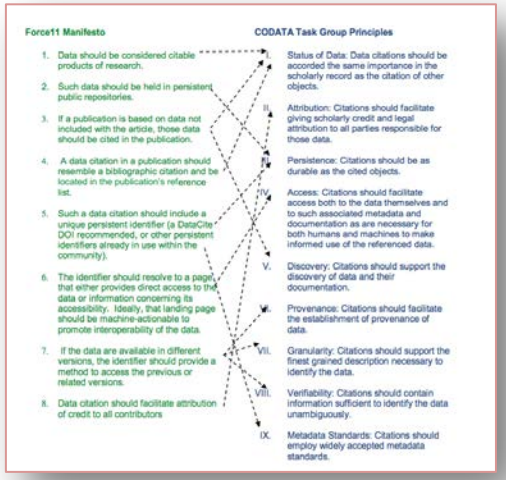
Joint Declaration of Data Citation Principles

(Overview)

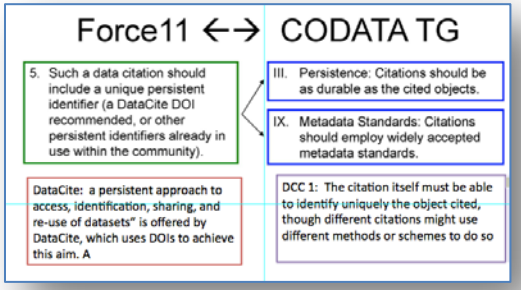
http://www.force11.org/datacitationsynthesisgroup

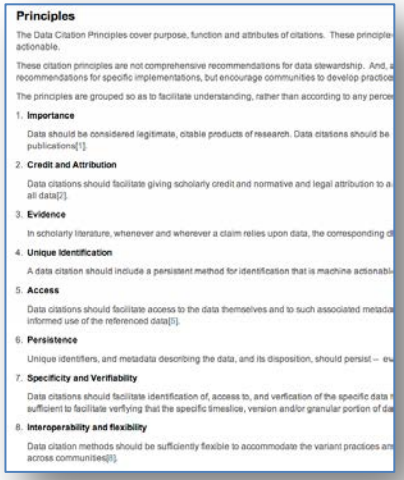
The Data Citation Synthesis Group

Background



Process





Synthesis → Community feedback → Revision → Dissemination

July-Sept 2013 | Nov-Dec 2013 | Jan 2014 | Now

Data Citation Principles: Open for Endorsement

Growing Adoption

Program on Information Science | MIT Libraries

https://www.force11.org/datacitation/endorsements

Joint Declaration of Data Citation Principles

Significance & Scope

- Sound, reproducible scholarship rests upon a foundation of robust, accessible data.
- Data should be considered legitimate, citable products of research.
- Data citation, like the citation of other evidence and sources, is good research practice.
- The Joint Principles cover purpose, function and attributes of citations.
- Specific practices vary across communities and technologies – we recommend communities develop practices for machine and human citations consistent with these general principles.

The Noble Eight-Fold Path to Citing Data

1. Importance
2. Credit and attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence
7. Specificity and verifiability
8. Interoperability and flexibility

Principles are supplemented with a glossary, references and examples
http://force11.org/datacitation

Purpose

1. **Importance.** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications [1].

2. **Credit and attribution**: Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data [2].

3. **Evidence.** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited [3].

Function

4. Unique Identification. A data citation should include a persistent method for identification that is machine-actionable, globally unique, and widely used by a community [4].

5. Access. Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data [5].

Attributes

6. **Persistence.** Unique identifiers, and metadata describing the data and its disposition, should persist -- even beyond the lifespan of the data they describe [6].
7. **Specificity and verifiability.** Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited [7].
8. **Interoperability and flexibility.** Data 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 citation practices across communities [8].

An Example

Placement of Citations

Intra-work:

- *Should provide sufficient information to identify cited data reference within included reference list.*
- *Citation to data should be in close proximity to claims relying on data.* ***[Principle 3]***
- *May include additional information identifying specific portion of data related supporting that claim.* **[Principle 7]**

Example*:* The plots shown in Figure X show the distribution of selected measures from the main data [Author(s), Year, portion or subset used].

Full Citation:

Citation may vary in style, but should be included in the full reference list along with citations to other types works.

Example:

References Section

Author(s), Year, Article Title, Journal, Publisher, DOI.

Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier.

Author(s), Year, Book Title, Publisher, ISBN.

Generic Data Citation

(as it appears in printed reference list)

Principle 2: Credit and Attribution (e.g. authors, repositories or other distributors and contributors)

Principle 4: Unique Identifier (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:** A persistent identifier that provides access and metadata

Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier

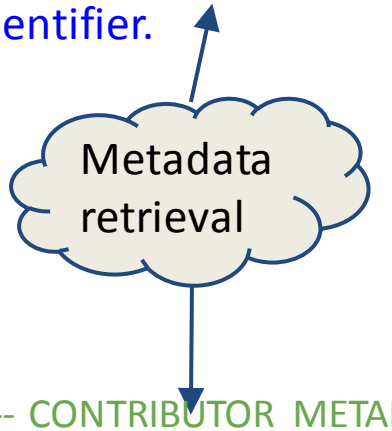
Principle 7: Specificity and verification (e.g. the specific version used).
Versioning or timeslice information should be supplied with any updated or dynamic dataset.

Note:

- Neither the format nor specific required elements are intended to be defined with this example. Formats, optional elements, and required elements will vary across publishers and communities. **[Principle 8: Interoperability and flexibility]**.
- As illustrated in the previous examples, intra-work citations may be accompanied with information including the specific portion used. **[Principles 7,8]**.
- As illustrated in the next example, printed citations should be accompanied by metadata that support credit, attribution, specificity, and verification. **[Principles 2, 5 and 7].**

Citation Metadata

Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier.




```
<!--- CONTRIBUTOR METADATA -->
<contributor role="
ORCIDid=">Name</contributor>

<!-- FIXITY and PROVENANCE --
<fixity type="MD5">XXXX</fixity>
<fixity type="UNF">UNF:XXXX</fixity>

<!-- MACHINE UNDERSTANDABILITY --
>
<content type>data</content type>
<format>HDF5</format>
```

EXAMPLE METADATA

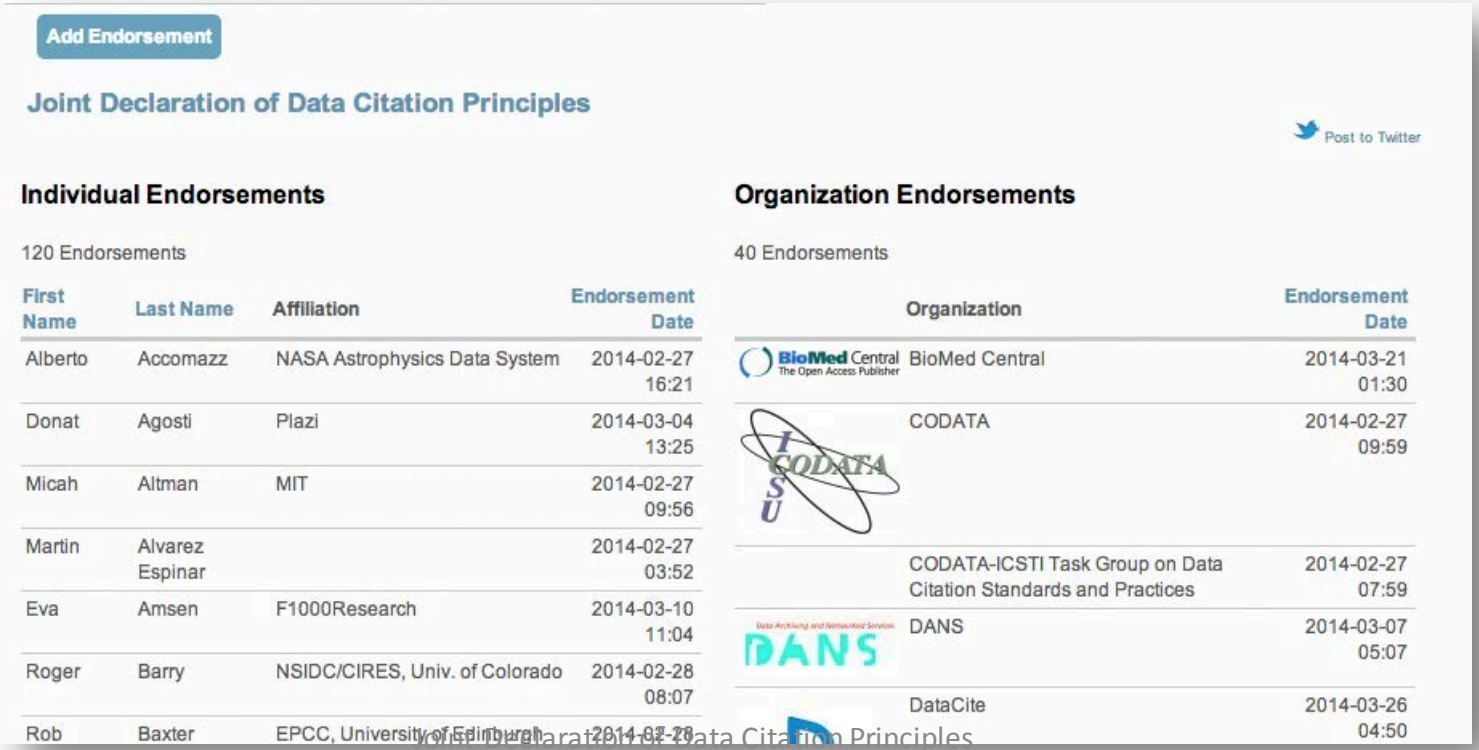
Note:

- Metadata location, formats, and elements will vary across publishers and communities. **[Principle 8]**
- Citation metadata is needed in addition to the information in the printed citation.
- Metadata describing the data and its disposition should persist beyond the lifespan of the data. **[Principle 6]**
- Citation metadata should support attribution and credit **[Principle 2]**; machine use **[Principle 5];** specificity and verification **[principle 7]**
- For example, additional citation metadata may be embedded in the citing document; attached to the persistent identifier for the citation, through its resolution service; stored in a separate community indexing service (e.g. DataCite, CrossRef); or provided in a machine-readable way through the surrogate ("landing page") presented by the repository to which the identifier is resolved.

For more detail, see the **References** section.
http://www.force11.org/node/4772

Endorse the Principles!

- http://www.force11.org/datacitation/endorsements



Add Endorsement

Joint Declaration of Data Citation Principles

Post to Twitter

Individual Endorsements

120 Endorsements

First Name	Last Name	Affiliation	Endorsement Date
Alberto	Accomazz	NASA Astrophysics Data System	2014-02-27 16:21
Donat	Agosti	Plazi	2014-03-04 13:25
Micah	Altman	MIT	2014-02-27 09:56
Martin	Alvarez Espinar		2014-02-27 03:52
Eva	Amsen	F1000Research	2014-03-10 11:04
Roger	Barry	NSIDC/CIRES, Univ. of Colorado	2014-02-28 08:07
Rob	Baxter	EPCC, University of Edinburgh	2014-02-28

Organization Endorsements

40 Endorsements

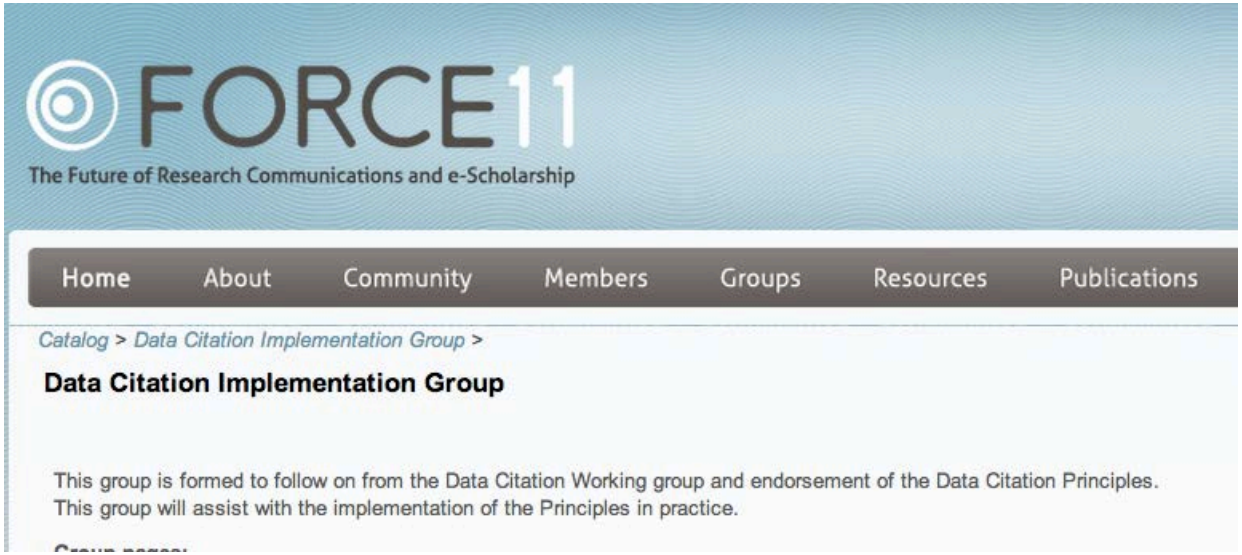
Organization	Endorsement Date
BioMed Central	2014-03-21 01:30
CODATA	2014-02-27 09:59
CODATA-ICSTI Task Group on Data Citation Standards and Practices	2014-02-27 07:59
DANS	2014-03-07 05:07
DataCite	2014-03-26 04:50

Joint Declaration of Data Citation Principles (Overview)

Join the Implementation Effort

- Implementation

  http://www.force11.org/node/4849



Notes & References

Notes

[1] CODATA 2013: sec 3.2.1; Uhlir (ed.) 2012, ch 14; Altman & King 2007

[2] CODATA 2013, Sec 3.2; 7.2.3; Uhlir (ed.) 2012,ch. 14

[3] CODATA 2013, Sec 3.1; 7.2.3; Uhlir (ed.) 2012, ch. 14

[4] Altman-King 2007; CODATA 2013, Sec 3.2.3, Ch. 5; Ball & Duke 2012

[5] CODATA 2013, Sec 3.2.4, 3.2.5, 3.2.8

[6] Altman-King 2007; Ball & Duke 2012; CODATA 2013, Sec 3.2.2

[7] Altman-King 2007; CODATA 2013, Sec 3.2.7, 3.2.8

[8] CODATA 2013, Sec 3.2.10

References

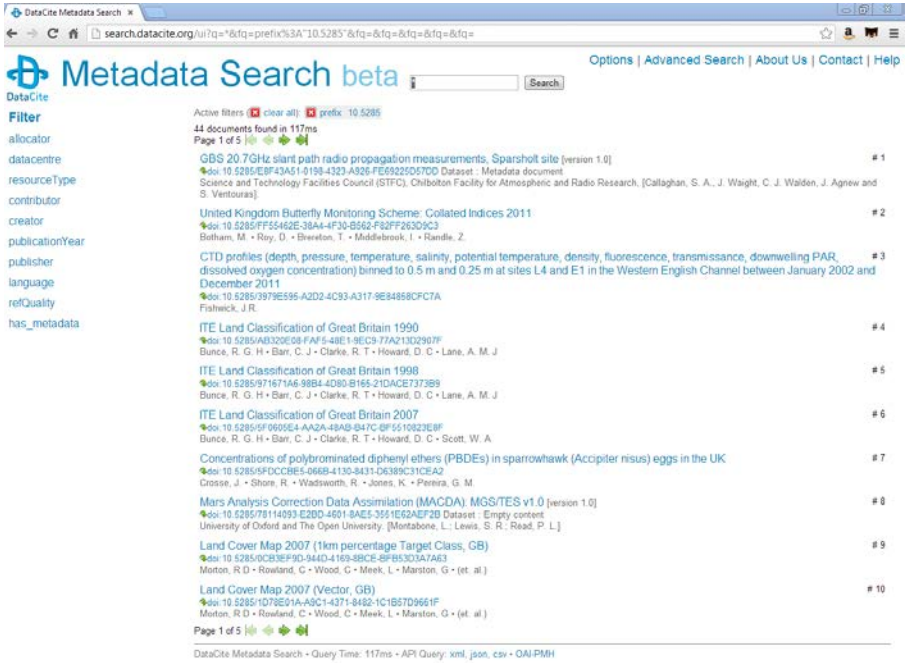
- M. Altman & G. King, 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data, *D-Lib*
- Ball, A., Duke, M. (2012). 'Data Citation and Linking'. DCC Briefing Papers. Edinburgh: Digital Curation Centre.
- CODATA-ICSTI Task Group on Data Citation, 2013; Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*
- P. Uhlir (ed.),2011. For Attribution -- Developing Data Attribution and Citation Practices and Standards. National Academies of Sciences

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



NERC's guidance on citing data and assigning DOIs can be found at: http://www.nerc.ac.uk/research/sites/data/doi.asp

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

Dataset has to be:

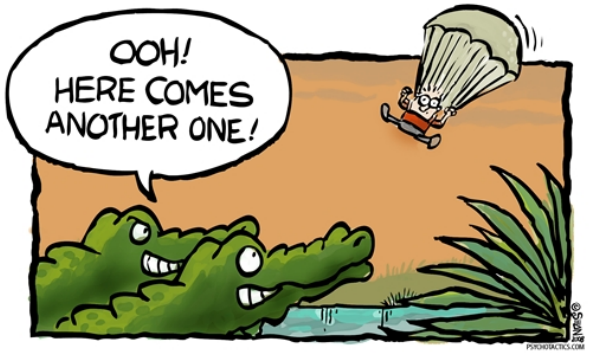
- Stable (i.e. not going to be modified)
- Complete (i.e. not going to be updated)
- Permanent – by assigning a DOI we're committing to make the dataset available for posterity
- Good quality – by assigning a DOI we're giving it our data centre stamp of approval, saying that it's complete and all the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

When a dataset is cited that means:

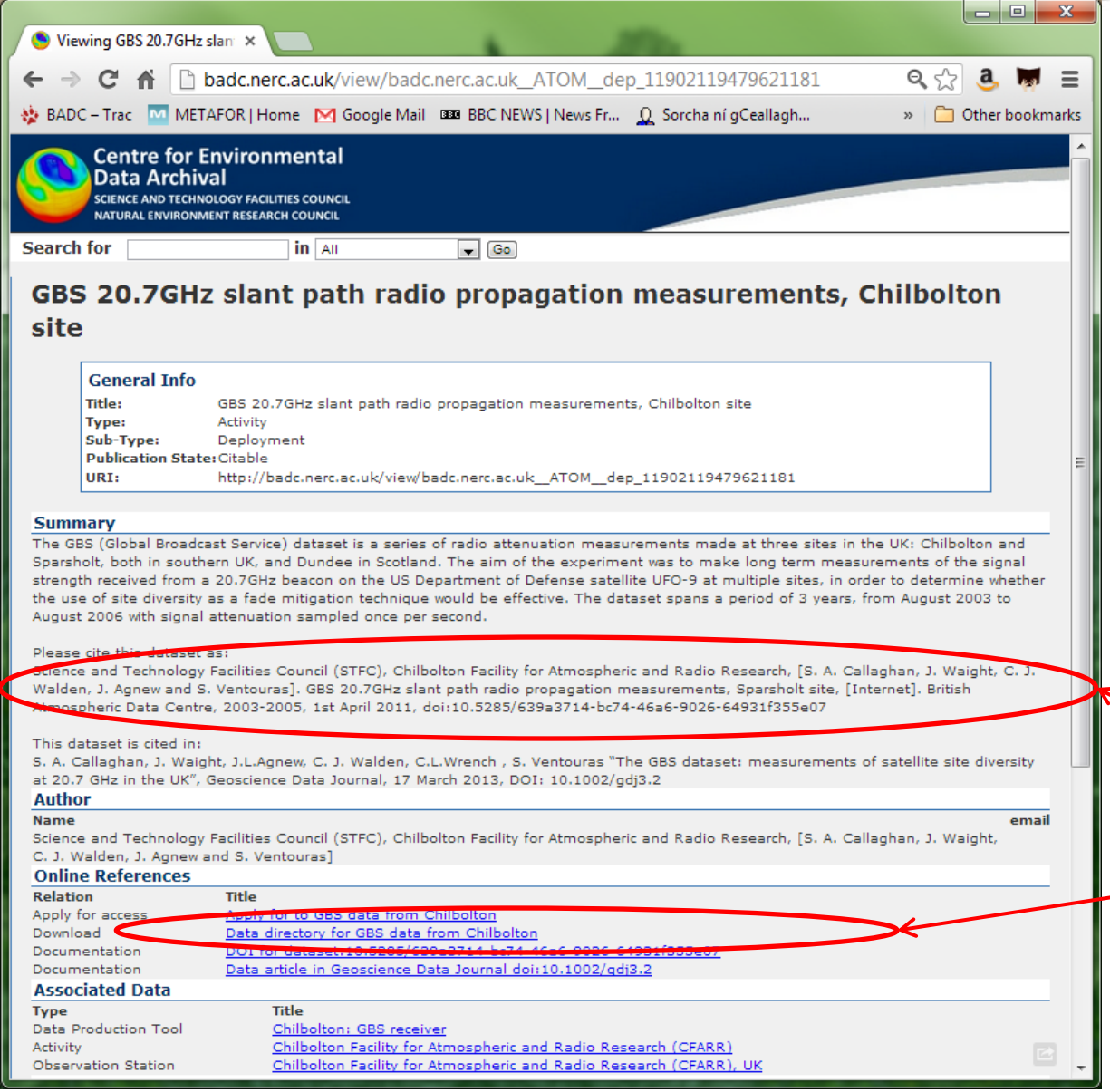
- There will be bitwise fixity
- With no additions or deletions of files
- No changes to the directory structure in the dataset "bundle"

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.



BAD LANDING PAGES



Dataset catalogue page (and DOI landing page)

Dataset citation

Clickable link to Dataset in the archive

ScienceDirect Journals | Books

Shopping cart | Remote access

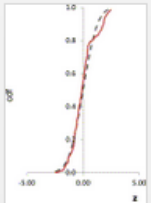
Download PDF | Export | More options... | Search ScienceDirect | Advanced search

Article outline | Show full outline

- Highlights
- Abstract
- Keywords
- 1. Introduction
- 2. Methods
- 3. Results and discussion
- 4. Conclusions
- Acknowledgements
- Annex 1. Full source information for r…
- References

Figures and tables

- Table 1
- Table 2
- Table 3



Learn more about our redesign on our blog. Click here for details.

ELSEVIER

Journal of Environmental Radioactivity

Volume 126, December 2013, Pages 314–325

Evaluating summarised radionuclide concentration ratio datasets for wildlife ☆

M.D. Wood[a], N.A. Beresford[b], B.J. Howard[b], D. Copplestone[c]

Show more

http://dx.doi.org/10.1016/j.jenvrad.2013.07.022

Open Access

Highlights

- The approac
- In contrast, t
- We propose
- Available da
- Generic CR

References

Albrecht et al., 2007 J. Albrecht, M. Abalos, T.M. Rice

Heavy metal levels in ribbon snakes (*Thamnophis sauritus*) and anuran larvae from the Mobile-Tensaw River Delta, Alabama, USA

Arch. Environ. Contam. Toxicol., 53 (4) (2007), pp. 647–654

View Record in Scopus | **Full Text** via CrossRef | Cited By in Scopus (4)

Barnett et al., 2013a C.L. Barnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone

Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in north-west England

NERC-Environmental Inf. Data Centre (2013) http://dx.doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196ece9ea

Barnett et al., 2013b C.L. Barnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone

Transfer parameters for ICRP reference animals and plants collected from a forest ecosystem

Radiat. Environ. Biophys. (2013) (in press)

Another example of a cited dataset

British Atmospheric Data Centre

NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE

NATURAL ENVIRONMENT RESEARCH COUNCIL

Data Citation and Data Publication

People have a good understanding of what a book or a journal article is:

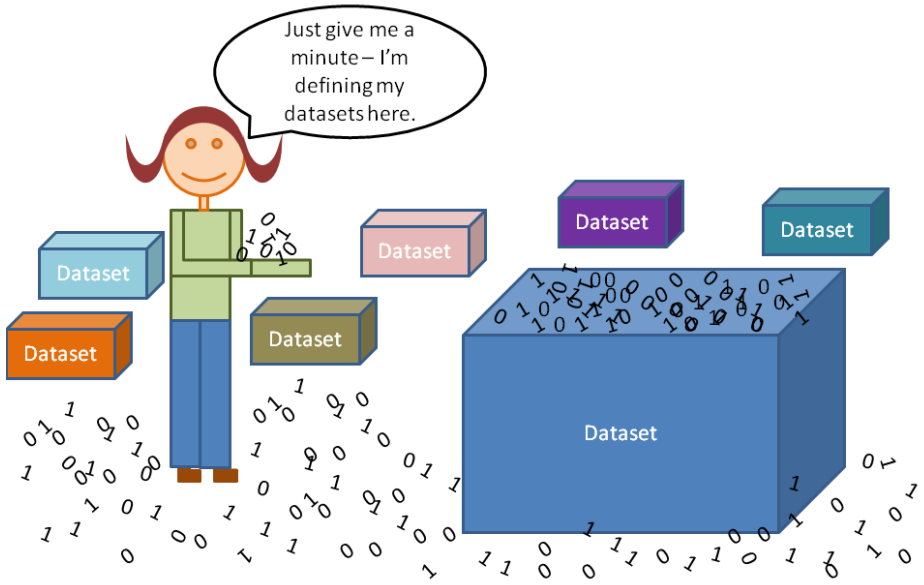
- clearly defined boundaries of what is and isn't in the book/article.

Datasets are more uncertain.

Dynamic data make it difficult to draw the boundaries of what is in the dataset and what isn't.

But to publish data correctly, we need to define and identify the datasets.

Data citation (and DOIs) help us do this!

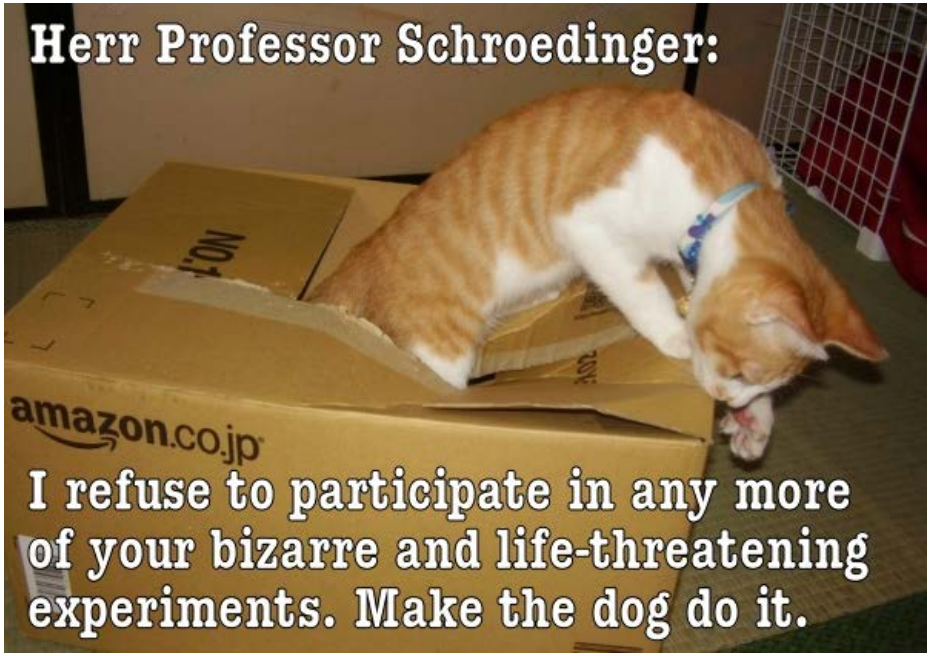


Data, Reproducibility and Science

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!



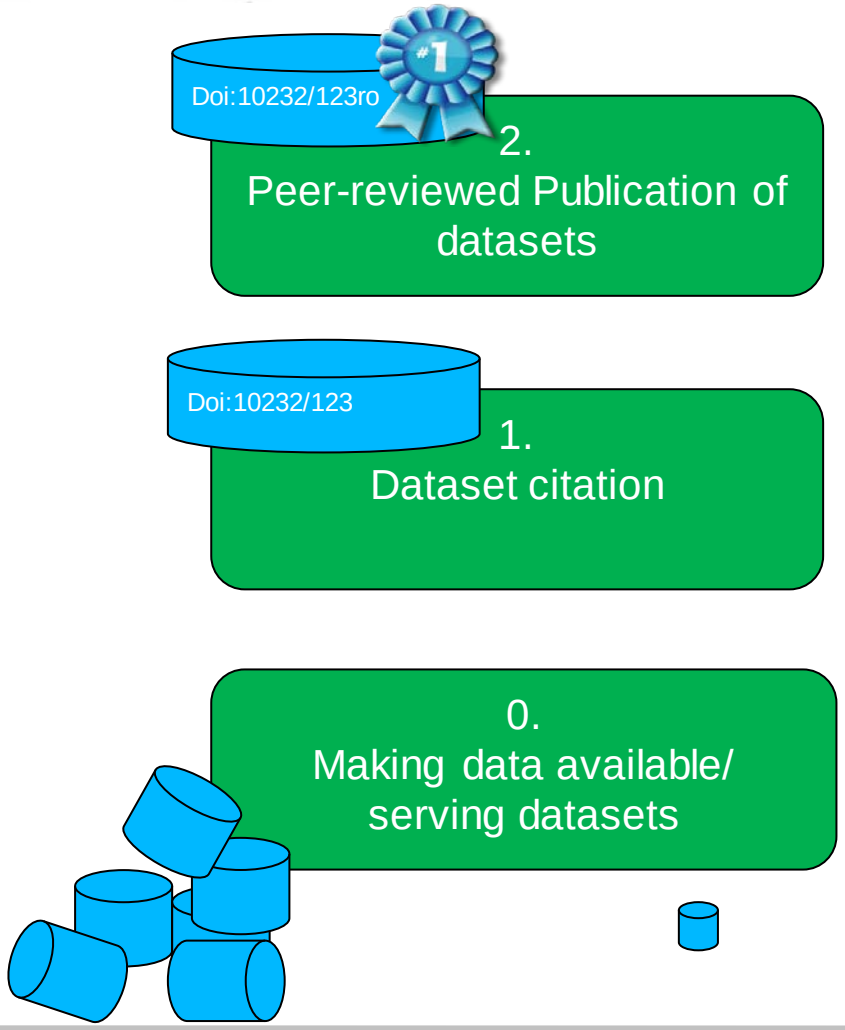
http://www.flickr.com/photos/31333486@N00/1893012324/sizes/o/in/photostream/

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/

Levels in data publication



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.

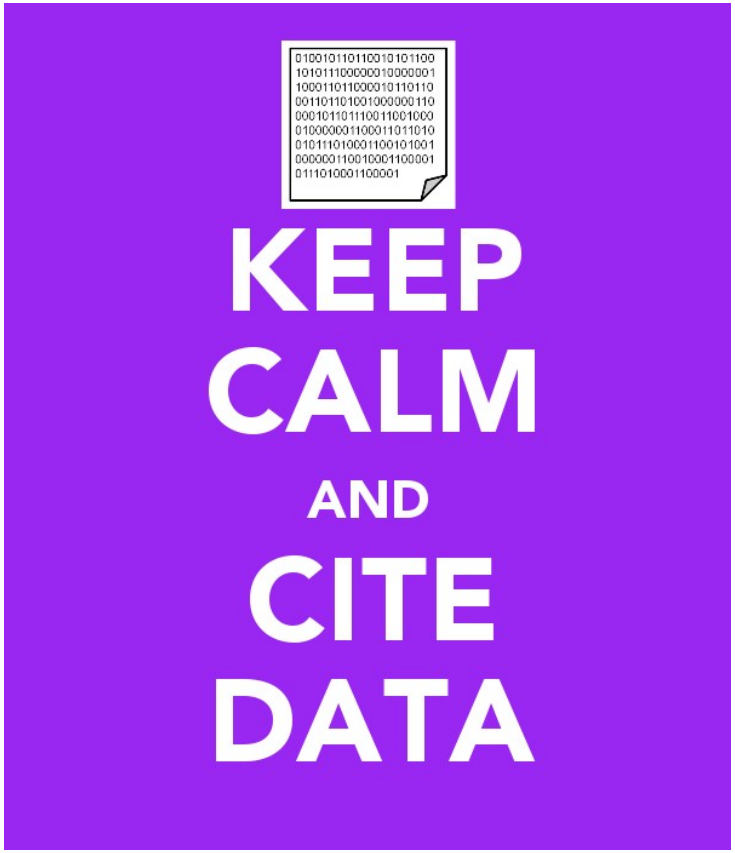
Citation is where we define what is in the dataset and what isn't. It needs identifiers that are permanent and unambiguous. Citing something means that you want to get the same thing back when you de-reference the citation - which is why we're using DOIs.

This is what data centres do as our day job – take in data supplied by scientists and make it available to other interested parties.
Data can and does change – "working data"

Conclusions

- Data citation makes us think more about our data
  - what is and what isn't part of the dataset
  - who is responsible for it/gets credit
  - how to identify it
  - what its purpose, function and attributes are
- Defining these things makes it easier for us and others to use the data
  - and facilitates data publication.

- Endorse the Joint Declaration of Data Citation Principles!
  - http://www.force11.org/datacitation/endorsements
- Join the implementation effort!
  - http://www.force11.org/node/4849



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

"Publishing research without data is simply advertising, not science" - Graham Steel

http://blog.okfn.org/2013/09/03/publishing-research-without-data-is-simply-advertising-not-science/

Thanks!

Any questions?

sarah.callaghan@stfc.ac.uk
@sorcha_ni
http://citingbytes.blogspot.co.uk/

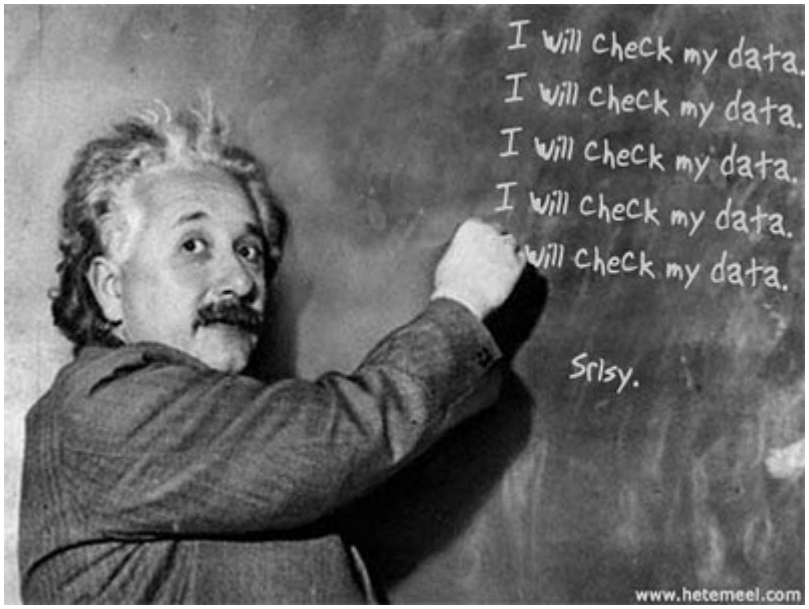


Image credit: Borepatch http://borepatch.blogspot.com/2010/06/its-not-what-you-dont-know-that-hurts.html