The International Journal of Digital Curation
Volume 7, Issue 1 | 2012

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 Environment 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 Moncoiffé,
British Oceanographic Data Centre

Kate Harrison, Ben Smith-Haddon, Anita Weatherby & Dan Wright,
Environmental Information Data Centre, Centre for Ecology and Hydrology

Abstract

The NERC Science Information Strategy Data Citation and Publication project aims to develop and formalise a method for formally citing and publishing the datasets stored in its environmental data centres. It is believed that this will act as an incentive for scientists, who often invest a great deal of effort in creating datasets, to submit their data to a suitable data repository where it can properly be archived and curated. Data citation and publication will also provide a mechanism for data producers to receive credit for their work, thereby encouraging them to share their data more freely.

International Journal of Digital Curation (2012), *7*(1), 107–113. http://dx.doi.org/10.2218/ijdc.v7i1.218

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

 doi:10.2218/ijdc.v7i1.218

Introduction

Through much of scientific history data has been a scarce resource, requiring significant efforts to obtain, but by contrast datasets were generally smaller and more easy to share in hard copy format, as either tables, pictures or graphs. As scientists' ability to collect more and increasingly detailed data has increased, their ability to publish it easily has decreased. Given that the currency of academic credit is based around the journal publication, and the historic difficulties associated with publishing data, it is not surprising that a scientific culture has arisen where data sharing is viewed with a variety of opinions from enthusiasm to skepticism or outright hostility. Knowledge is power, and in an increasingly competitive market for research funding sole possession of a significant dataset might be a key factor in ensuring continued funding.

The benefits of sharing data are many, including the ability to discover and reuse data which has already been collected, thus avoiding redundant data collection and saving time and money; and providing opportunities for collaboration. For this reason, research funders are keen to encourage data sharing. The tension on the researchers' side is that there is (currently) no universally accepted mechanism for data creators to obtain academic credit for their dataset creation efforts. Consequently, they often prefer to hold the data until they have extracted all the possible publication value they can. Though completely understandable, this behaviour comes at a cost for the wider scientific community.

A tension therefore exists between the need to share data to encourage reuse and collaboration, whilst still ensuring that the shared data is of good scientific quality and is suitable for reuse. In parallel to this is the data creator's need for attribution and credit, whilst they balance the reputational risks associated with sharing (including the discovery of errors in the data, increased opportunity for collaboration) versus the benefits of not sharing (such as maximising publications and research funding).

This paper details the work done by the NERC Science Information Strategy Project on Data Citation and Publication, and attempts to put the concepts of data citation and publication into the context of work done by the NERC-funded research community. The project is being run as a collaboration of the NERC environmental data centres, who wish to encourage researchers to deposit data in the archives where it can be curated and managed properly. Data citation and publication is being proposed as an incentive for researchers to do just this, and thereby avoiding the situation humorously outlined in Brown (2010).

"Publishing" Versus "publishing"

It is now possible to "publish" data relatively easily; at its most basic all a researcher has to do is to stick the files on a website somewhere. This makes the data open, but without any form of long-term commitment. There are no guarantees that the data will still be there in six months, or that the files won't get corrupted. Furthermore, it is possible that a scientist who isn't the data creator won't be able understand the contents or even open the files at all. Even if the dataset is readable and has sufficient metadata, there is no information about the scientific quality of the dataset, other than that attached to the creator's reputation.

By contrast, a formal "Publishing" process adds value to the dataset for the future consumers of the data. This may be by providing an indication of the scientific quality and importance of the dataset (as measured through a process of peer-review), or by ensuring that the dataset is complete, frozen, and has enough supporting metadata and other information to allow it to be used by others in the years to come. "Publishing" implies a commitment to persistence of the data. It also provides a mechanism for allowing data producers to obtain academic credit for their work in creating the datasets.

The notion of formally "Published" data does not necessarily imply that the data would be open, but there is no reason why "Published" data should not be open. Figure 1 gives a schematic example of this.

There have been many discussions held about closed versus open data, and there will be many more in the future. What is generally well agreed is that it is no longer appropriate to keep significant datasets stored on a single hard drive, or several CDs in a drawer in an office somewhere. The recent Climategate scandal showed that the general public do indeed have an interest in the work that their taxes are funding. The UK government also wish to make all data from publicly funded research available to the public for free.

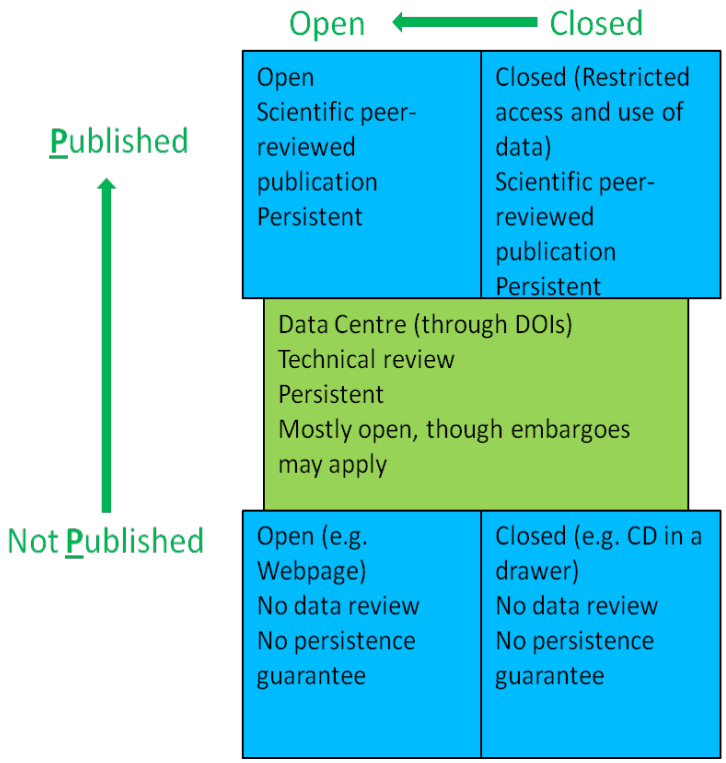

Figure 1. The tension between open and closed publication and Publication. (DOIs are digital object identifiers)

To a scientist, there is little benefit from making their dataset available as a free download from a webpage, unless they work in certain areas of science where this is expected. In fact, the reputational risk of doing so (particularly if others find errors, or

 doi:10.2218/ijdc.v7i1.218

worse, take advantage of the dataset to earn new research funding) and the extra work involved in doing so, might mean that the scientist would prefer to store the data on a closed server. 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.

Data Citation and Publication and the NERC Environment Data Centres

It is commonly accepted that data curation is a difficult job, and most data producing scientists have neither the time nor the inclination to focus on it. It is for this reason that NERC funds six data centres, which between them have responsibility for the long-term management of NERC's environmental data holdings. NERC researchers are expected to liaise with these data centres to determine how and what portions of their data should be archived and curated for the long term, and then work with data centre staff to ingest the dataset, together with its accompanying metadata and documentation, into the archives.

NERC are also keen to obtain good value from the research they fund and so have set up the Science Information Strategy (SIS) to provide the framework for NERC to work more closely and effectively with its scientific communities in delivering data and information management services.

The NERC SIS data citation and publication project aims to create a way of promoting access to data, while simultaneously providing the data creators with full academic credit for their efforts. The project also aims to implement a process to ensure the technical and scientific quality of the resulting datasets. To achieve this, we are developing a mechanism for the formal citation of datasets held in the NERC data centres, and are working with academic journal publishers to develop a method for the scientific peer-review and publication of datasets.

The first step in this project is to formalise a method for citing datasets, and to encourage the NERC scientific community to use it as standard when discussing datasets in the literature.

Citation of Data Using Digital Object Identifiers (DOIs)

Anyone can reference a dataset stored on the internet by using an appropriate form of words, plus a URL linking to the page where the dataset can be found. However, URLs are renowned for breaking, and so do not deliver the stability that one expects for a formal citation. It is for this reason that we have decided to use Digital Object Identifiers (DOIs) to signify datasets that are complete, in a useable format, stable (changes are implemented by publication of new versions), have valid metadata, have passed the quality control checks within the domain of expertise of the data centre, and have long-term stewardship guaranteed by that data centre, underwritten by the ICSU World Data System. This provides the basis for a dataset to be cited as if it were a research paper, putting it on a par with other scientific outputs.

The NERC data centres are not the only groups to use DOIs for citing datasets. For example, in the Earth Sciences, the Pangaea data archive[1] cite their datasets using DOIs, and the ISIS[2] pulsed neutron and muon source issues DOIs to their experiments. Scientists are already used to citing papers using DOIs, so it is only a small change to their behaviour to get them to cite data in the same way. Using DOIs for data also allows us to piggy-back on various pre-existing citation metrics, without having to invent new ones.

At the time of writing, the data citation project has successfully assigned DOIs to 14 datasets held in the NERC environmental data centres. We are still in a testing phase, and guidelines for what constitutes a dataset suitable for DOI assignment are in development. At the moment, all the DOIs that have been assigned are to completed, legacy datasets in the archive. We anticipate that in the near future, dataset authors will be creating datasets with the aim of getting a DOI for them when they're completed. Permission to assign a DOI (should the dataset meet the criteria) will be sought from the data authors as part of the creation of the data management plan. The technical criteria for DOI assignment will also be presented to the dataset author at this stage, allowing them to ensure that their data meets the criteria.

The DOI assignment account has been issued to NERC by the British Library, acting on behalf of DataCite,[3] as part of a pilot project by DataCite. NERC are not the only DOI-issuer for data in the UK. Other participants include the Archaeology Data Service, the UK Data Archive and the Bodleian Library at the University of Oxford. The methods proposed by the NERC data centres for the citation of their datasets could as easily be applied by any other data repositories, provided they met the DataCite criteria for being DOI minters.

It is anticipated that a data citation (with DOI) will be of value to the authors of the dataset, even if they never then go through the scientific peer-review process associated with journal publication. A helpful analogy would be to consider a dataset in a data centre (without DOI) as equivalent to grey literature, a dataset with a DOI citation as a paper published in conference proceedings, and a dataset published in a formal data journal as equivalent to an academic journal article.

Peer-Review of Data

As data centres, our main area of expertise is in data management, supported by general rather than specialised domain knowledge. Consequently, whilst we can make quality claims about the completeness of the metadata and appropriateness of file format with total confidence, we cannot make equivalent claims about the scientific validity of the dataset. This scientific quality must be established through a process of peer-review by other specialists in the data creator's field. Such processes are already well-established in traditional academic publishing and are well understood by scientists. We are therefore currently working with publishers to create a mechanism for the scientific publication of datasets that includes full peer review. This process is still in early stages, as most of the project effort has concentrated on establishing the

[1] Pangaea Data Archive: http://pangaea.de

[2] ISIS: http://www.isis.stfc.ac.uk/

[3] DataCite: http://www.datacite.org

 doi:10.2218/ijdc.v7i1.218

mechanisms for citation of data that will be required before formal publication can be established. Some work on what a reviewer might look at when reviewing a dataset has been done in the context of the CLADDIER project (Lawrence et al., 2011).

It is worth noting that peer-review shouldn't be a pre-requisite for informal data sharing between scientists. As well as our data curation activities, the NERC data centres act as facilitators for the exchange of data, and datasets can be shared at any point during the ingestion process, should the authors so wish. As DOIs can only be assigned once the dataset is complete and frozen, there is plenty of time before a DOI is assigned when informal sharing is the only option.

Publication of Datasets

The mechanism proposed for the scientific peer-review and formal Publication of a dataset rests on top of the mechanism for citation (see figure 2). Our current plans for data Publication involve working with academic publishers to develop a new style of article: a data paper, which would describe the dataset, providing information on the what, where, why, how and who of the data. The data paper would contain a link back (a DOI) to the dataset in its repository, and the journal publishers would not actually host the data. This means that even in situations where the data paper might be restricted access, the dataset could still be open.

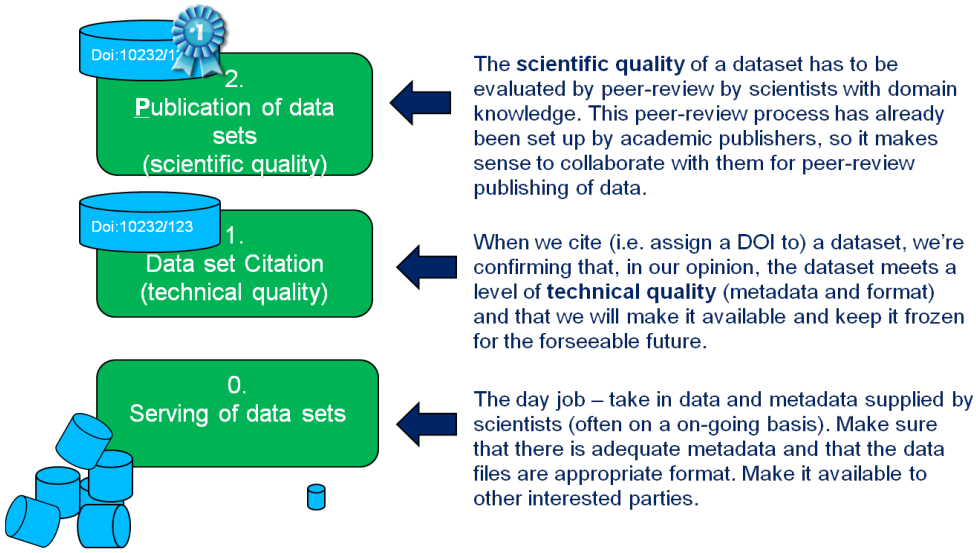

Figure 2. Relationship between dataset serving, citation and Publication.

This parallels with the well-established practise in astronomy, climate science and other fields of the "data release paper", which acts as a proxy for the dataset and describes the technical form and scientific content of the dataset, and acts as a guide to its use for other researchers. Those using the dataset can then reference the proxy paper, often generating large citation lists and providing recognition to the researchers who generate them. The mechanism outlined earlier for data citation allows the datasets to be citeable without the need for a "data release paper" or a proxy paper, making it quicker and easier for datasets to be cited.

Formal data journals already exist. For example, Earth System Science Data[4] publishes the data papers associated with datasets stored in other repositories, while Geochemistry, Geophysics, Geosystems (G3)[5] publishes data briefs.It is worth noting that this method of peer-review and Publication of data is not definitive. For example, the Planetary Data System (PDS)[6] of the Jet Propulsion Laboratory has extensive experience in the peer review of scientific data products, as well as publication and citation of scientific data products.

Conclusions

Data citation and publication will ensure that data will be considered as a first class research output that will be available, peer-reviewed, citable, easily discoverable and reusable. The mechanisms for citation and publication will facilitate data transparency and scrutiny, and will be used by researchers to increase their academic status, thereby providing an incentive for them to archive and document their data appropriately. This will result in significant gains for both the current research community and scientists for decades to come.

Acknowledgements

The work described in this paper is funded by NERC as part of its Science Implementation Strategy Programme.

References

Lawrence, B., Jones, C., Matthews, B., Pepler, S. & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. International Journal of Digital Curation, 6(2). Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/181

Brown, C.T. (2010). My Data Management Plan: A satire in blog. Retrieved from http://ivory.idyll.org/blog/may-10/data-management.html

[4] Earth System Science Data: http://earth-system-science-data.net/

[5] G3: http://www.agu.org/journals/gc/

[6] PDS: http://pds.nasa.gov/