ARIADNE

Section 1: Details Text Fields

1.Title of Article

Enter in box below (try if possible to avoid exceeding 8 words):

Data Citation and Publication by NERC's Environmental Data Centres

2. Article byline - enter in box below

Author Name(s) and by-line in which the author(s) outline(s) in up to 25 words the content of the article; for example:

e.g.

June Brown describes new models of working to assist in the implementation of Dublin Core metadata.

Sarah Callaghan, Roy Lowry, David Walton and members of the Natural Environment Research Council Science Information Strategy Data Citation and Publication Project team describe their work in NERC's Environmental Data Centres.

Section 2: Article Text

Your article goes beneath **Introduction***… try to use some sub-headings in your text, suggestions follow. Article length: c. 2300 – 4000, though do not worry if you exceed the maximum by a small degree, better to submit on time than worry reducing word total needlessly.*

Why Cite and Publish Data?

Data are the foundation upon which scientific progress rests. Historically speaking, data were a scarce resource, but one which was (relatively) easy to publish in hard copy, as tables or graphs in journal papers. With modern scientific methods, and the increased ease in collecting and analysing vast quantities of data, there arises a corresponding difficulty in publishing this data in a form that can be considered part

of the scientific record. It is easy enough to ‘publish’ the data to a Web site, but as anyone who has followed a broken link knows, there is no guarantee that the data will still be in place, or will not have changed, since it was first put online. A crucial part of science is the notion of reproducibility; if a dataset is used to draw important conclusions, and then the dataset changes, those conclusions can no longer be re-validated by someone else.

Data curation is a difficult and time-consuming job, and most scientific data producers have neither the time, funding, nor inclination to do it. It makes sense to take advantage of economies of scale to support the expertise required for data curation, and for this reason the Natural Environment Research Council (NERC) funds six data centres (see Table 1) which among them have responsibility for the long-term management of NERC's environmental data holdings. Researchers receiving funding from NERC, both in research centres and in HEIs, are expected to liaise with the appropriate data centre to determine how and what portions of their data should be archived and curated for the long term.

Data Centre	**Area of Interest**
British Atmospheric Data Centre (BADC)	Atmospheric science
National Geoscience Data Centre (NGDC)	Earth sciences
NERC Earth Observation Data Centre (NEODC)	Earth observation
British Oceanographic Data Centre (BODC)	Marine Science
Polar Data Centre (PDC)	Polar Science
Environmental Information Data Centre (EIDC)	Terrestrial and freshwater science, Hydrology and Bioinformatics

Table 1: List of the NERC Environmental Data Centres and their scientific areas of interest

Even when most of the effort of data curation is carried out by the data centres, there is still significant amounts of work that must be done by the researchers before the datasets can be archived properly. For example, the data must be documented in such a way that future users can understand what is measured in the datasets and have the supporting information about things like instrument calibration and location, times of measurements etc. Data submitted to a file-based archive should be in file formats that are standard for the community and non-proprietary.

NERC and their environmental data centres want to ensure that the archived datasets are first-class scientific objects, and that the researchers responsible for creating them receive appropriate recognition for their efforts. NERC 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. We are therefore 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 peer-review and formal publication of datasets.

Unsurprisingly, this process is still ongoing. This article documents the path we took and the decisions we made in order to implement a method of data citation. By no means is it promoted as the optimum path, but instead is presented in order to allow others to see the potholes we encountered along the way.

Previous Projects

The NERC Data Citation and Publication Project began in April 2010, but before that members of the project team were involved in several initiatives looking at data citation and publication.

CLADDIER

The Citation, Location, And Deposition In Discipline & Institutional Repositories (CLADDIER) was funded by JISC under the Call for Projects in Digital Repositories (March 2005). Its aim:

> *'The result will be a step on the road to a situation where active environmental scientists will to be able to move seamlessly from information discovery (location), through acquisition to deposition of new material, with all the digital objects correctly identified and cited.'*

It did a lot of thinking about the roles, terminology, processes, etc, involved in data publication, and produced a method for writing the citation of a dataset, equivalent to how one would cite a journal paper. Use of this suggested citation structure was implemented in the British Atmospheric Data Centre (BADC) where dataset catalogue pages gave a recommended citation. For example:

> 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.

From what we can tell from Internet searches such as Google Scholar, this form of citation does not appear to have been adopted to any great extent by the scientific community.

Lawrence et al, 2011 [1] provide a summary about the work done on data citation and publication in the CLADDIER Project.

OJIMS

The Overlay Journal Infrastructure for Meteorological Sciences (OJIMS) Project was a follow-on project from CLADDIER and was also funded by JISC and NERC. It investigated formal journal publication of data and produced a demonstration data journal, which used overlay mechanics to create data description documents describing a dataset. Much like CLADDIER, it did not examine the mechanics of

linking (using URIs) in any depth, instead focusing on the mechanics of the overlay documents.

OJIMS also constructed and evaluated business models for potential overlay journals, and surveyed the user community in the meteorological sciences to determine their opinions of data publication and data repositories. Further information can be found in *Ariadne* [2][3].

SCOR/IODE/MBLWHOI Library Data Publication Working Group

This is a working group initially constituted in 2008 by the Scientific Committee on Oceanic Research (SCOR), an international non-governmental organisation promoting international scientific collaboration, and International Oceanographic Data and Information Exchange (IODE), an inter-governmental programme to promote sharing of oceanographic data. Parallel activities in the Woods Hole Library were identified which resulted in their joining the group in 2009. The group is currently chaired by the British Oceanographic Data Centre (BODC) (Roy Lowry) but will be jointly chaired by SCOR (Ed Urban) and MBLWHOI Library (Lisa Raymond) from 2012.

The group was set up primarily to engage the IODE national data centres in data publication, thereby providing parallel infrastructure to the German Pangaea data centre. There have been four meetings to date plus a related meeting of the parallel activity at Woods Hole:

- Ostend, June 2008 [4]
- Ostend, March 2009 [5]
- Jewett Foundation Woods Hole Data Repository Project Meeting, WHOI, April 2009
- Paris, April 2010 [6]
- Liverpool, November 2011 [7]

Data Publication Issues for Data Centres

The result of this work has been the recognition of a number of data publication issues which data centres will need to address:

Data centres regard datasets as dynamic concepts, whereas publishers regard a dataset as a static concept. Whilst it might be possible to cite a dynamic entity, this has little value in the scientific context where the citation is a proxy for an instance of content.

IODE data centres conform to a model where the data are manipulated and extended (eg metadata creation) to enhance their value with particular reference to increasing the fitness of the data for future recycling. This is a continual process largely based on changes to resources that are shared across the centre's entire data system. At any moment in time, dynamic datasets are assembled and served as the 'best currently available' version of the data. This sits uncomfortably with a publication model based on versioning where all versions of a particular dataset are available on request.

In Paris, a decision was taken to contact IODE data centres directly to get them involved in data publication. This resulted in an enthusiastic response. Everybody

wants to be involved in data publication. The trouble is that nobody seems to know where to start. In an attempt to break this *impasse*, BODC engaged in 'trailblazing' pilot project work that was reported to the CODATA conference at the end of October 2010 and at two IODE meetings in Liege in March 2011. The work has also become the focus of BODC involvement in the NERC SIS Data Citation and Publication Project.

BODC Pilot Projects

At BODC, Roy Lowry, Gwen Moncoiffe and Adam Leadbetter have been looking at how to map activities in the data centre to the data publishing paradigm. The approach has been to identify data that could be tagged by permanent identifiers such as DOIs and subsequently made available as static digital objects. This has led to the concept of the BODC Published Data Library, which began development in the latter half of 2011.

The Library comprises a catalogue of snapshot copies of datasets from one of two sources. The first source is exports of specified collections of related data that have been ingested into the BODC system. This reproduces BODC's data publication activities in the 1990s which created project datasets on CD-ROM and consequently has been dubbed '21st Century CD-ROMs'. The data from the Marine and Freshwater Microbial Biodiversity (M&FMB) Project together with a number of specific cruise datasets have been identified for the initial trials of the Library.

The second data source is data that have not been ingested by BODC, but are destined to be ingested in the future. Providing these data satisfy the basic criteria for publication, and are supplied to technical standards (use of long-lived formats and adequate labelling of the data streams) specified by BODC - with accompanying metadata deemed satisfactory by BODC - they will then be published immediately and ingested later. This provides a route for scientists requiring citations sooner rather than later.

The Published Data Library publication procedure is as follows:

- Obtain a DOI through the NERC arrangement with the British Library
- Prepare a catalogue entry and landing page in the PDL area on the BODC Web site
- Store a copy in a suitable repository that guarantees the dataset will be available and unchanged for the foreseeable future. In the short term, the BODC Data Vault will be used, but the project will also trial the IODE Published Ocean Data D-Space repository.
- Post a NERC metadata (ISO19115 profile encoded according to ISO19139 schema) record that covers the dataset (but may also cover other related datasets)
- Extract a Dublin Core record from this NERC metadata record.

Citation of Datasets

At its most basic level (the bottom layer in Figure 1), the main job of a data centre is to 'serve' data to its user community, ie take in data supplied by scientists and make them available to other users. Datasets may not be clearly defined, and will be served

according to the most recent data in the archive. Previous versions of datasets may not be kept, and no guarantees are made about the completeness or stability of the dataset. This is business as usual for the data centres.

The top level of Figure 1 shows what the project is aiming for, i.e. a formalised method of peer-reviewing and peer-approving datasets, of the sort that is traditionally associated with scientific journal publication. This provides the dataset with a 'stamp of approval' and provides data producers with academic credit, encouraging them to ensure their data are of good (scientific) quality (which is checked via the process of peer review [8]) and that their data are stored in a trusted [9] data repository.

Note that we draw a clear distinction between publishing or serving data, ie making data available for consumption (for example on the Web), and Publishing (note the capital 'P'), which is publishing after some formal process which adds value for the consumer; for example, a PloS ONE type review, or an EGU journal type public review, or a more traditional peer review *and* which provides commitment to persistence of the dataset being Published.

In order to bridge the gap between formal Publication of data and simple serving of data, a method of citing datasets is required. This citation provides a bridge between data and other publications, and is a useful object in its own right, as well as providing an essential step on the road to data Publication. At this time, the NERC Data Citation and Publication Project is primarily concentrating on formalising a method of citing the datasets held in the NERC data centre archives.

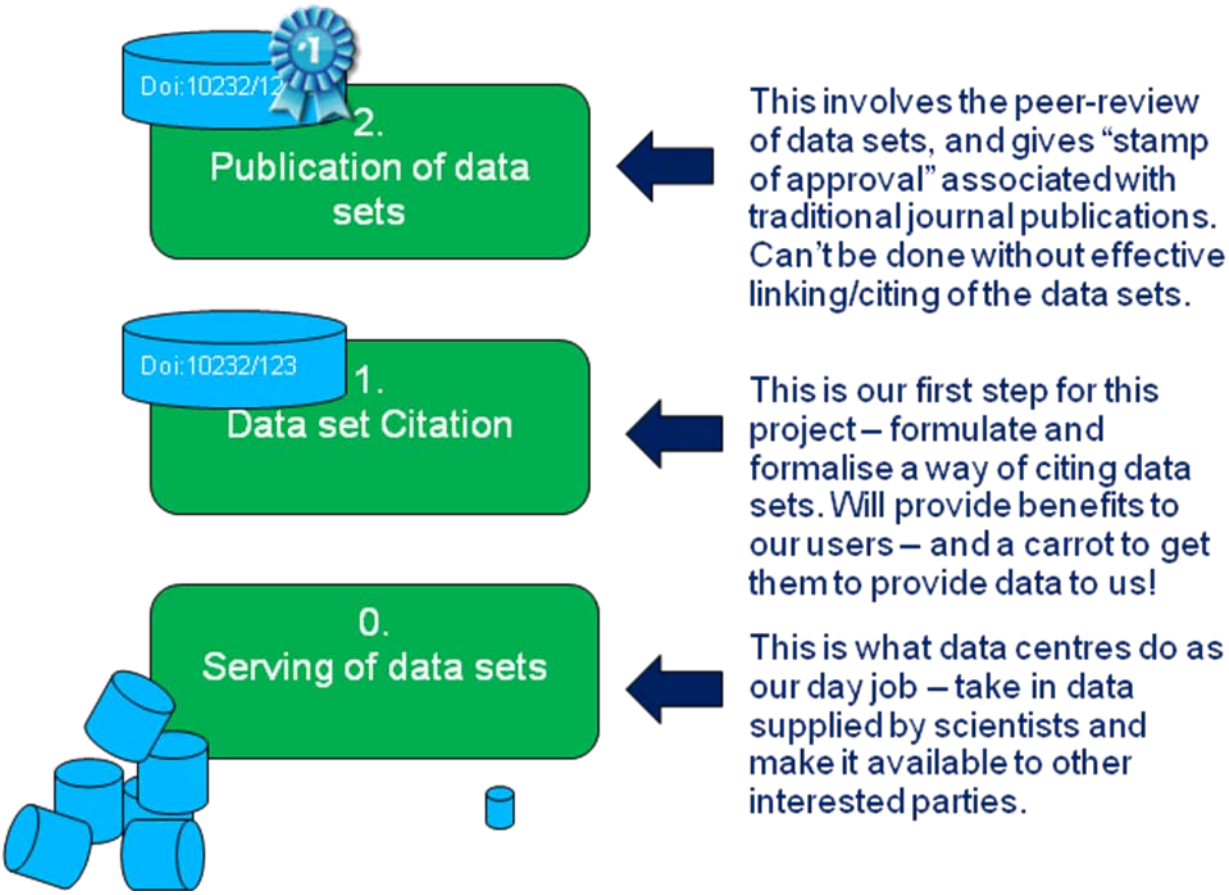

Figure 1: Serving, citing and publishing data

Citation Mechanism

As mentioned in the previous section describing the work done by the CLADDIER project, the BADC has provided a recommended citation for each of its datasets, which includes a URL link to the dataset catalogue page. These citations are not being used; we suspect in part because URLs are not trusted as a persistent link.

For this reason, we took the decision to use digital object identifiers (DOIs) to link to and cite our datasets because:

- They are actionable, interoperable, persistent links for (digital) objects
- Scientists are already accustomed to citing papers using DOIs, and so are familiar with and trust them.
- Pangaea [10] assign DOIs, and Earth System Science Data (ESSD) [11] use DOIs to link to the datasets they publish
- The British Library and DataCite gave us an allocation of 500 DOIs to assign to datasets as we saw fit.

The British Library (acting on behalf of DataCite) set NERC up with a DOI minting account which covers all the datasets assigned DOIs in all NERC data centres.

DOIs all follow the same format; a prefix (in NERC's case 10.5285) followed by a unique string of NERC's choice. The project team decided to use GUIDs (Globally Unique Identifiers) as the unique string.

The value of a GUID is represented as a 32-character hexadecimal string, such as {21EC2020-3AEA-1069-A2DD-08002B30309D}, and is usually stored as a 128-bit integer. The total number of unique keys is 2^{128} or 3.4×10^{38} — roughly 2 trillion per cubic millimetre of the entire volume of the Earth. This number is so large that the probability of the same number being generated twice is extremely small.

The disadvantage is that GUIDs do not look attractive, and there is no NERC branding in the string.

The advantage is that the opaqueness makes them easily transferable between data centres (if needed), and researchers will not be tempted to type them in (risking typographical errors) but instead will copy and paste them.

What Makes a Data Set Citable?

Before assigning a DOI to a dataset, certain criteria must be met, namely that the cited dataset has to be:

- Stable: not subject to modification
- Complete: not subject to updates
- Permanent: by assigning a DOI, data centres are committing themselves to making the dataset available for the foreseeable future
- Of good (technical) quality: by assigning a DOI, data centres are giving it a 'stamp of approval', stating it is complete with all metadata made available

Technical Quality versus Scientific Quality of Datasets

When data centres cite (ie assign a DOI to) a dataset, they are confirming that, in our opinion, the dataset meets a level of *technical quality* (metadata and format) and that they will make it available and keep it frozen for the foreseeable future.

The *scientific quality* of a dataset has to be evaluated through 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 in peer-review publishing of data (see later in this article).

Knowing that a dataset meets a given level of technical quality will make the scientific review process easier for the reviewers, as this will have confirmed that: the dataset is in the right format; that files can be opened; that variables are meaningfully named, etc. The scientific reviewer can then focus on reviewing the dataset in terms of whether the data are scientifically useful and meaningful.

The objective of data management within a data centre is to ensure that data may be reused with confidence decades after their collection without the need for any kind of communication with the scientists who collected that data. The following technical criteria, based on good practice criteria adopted across the NERC Environmental Data Centres, must be met for a dataset to have a DOI assigned to it by NERC.

Requirements Related to Datasets

The format must be well-documented and conform to widely accepted standards, such as ASCII or NetCDF. Preferably, data formats should conform to internationally agreed content standards, such as CF-compliant NetCDF or SeaDataNet ASCII spreadsheet format.

The format must be readable by tools that are freely available now and, ideally, are likely to remain freely available indefinitely.

Data files should be named in a clear and consistent manner throughout the dataset with filenames (rather than pathnames) that reflect the content and which uniquely identify the file. Filename extensions should conform to appropriate extensions for the file type. Filenames should be constructed from lower case letters, numbers, dashes and underscores and be no longer than 64 bytes.

Parameters in data files should either be labelled using an internationally recognised standard vocabulary specifically designed for labelling parameters, such as the BODC Parameter Usage Vocabulary or CF Standard Names, or by local labels that are accompanied by clear, unambiguous plain-text descriptions.

Units of measure must be included for all parameters and labelled following accepted standards such as UDUNITS or the SeaDataNet units vocabulary.

Data must be accompanied by the following XML metadata documents. The first is a Dublin Core metadata record including the dc:title, dc:creator, dc:subject, dc:period, dc:description, dc:contributor, dc:date, dc:language and dc:coverage elements. The second is a discovery metadata record conforming to a recognised standard such as:

- European Directory of Marine Environmental Data (EDMED)

- Global Change Master Directory (GCMD) Directory Interchange Format (DIF)
- Marine Environmental Data and Information Network (MEDIN) ISO19139 discovery metadata profile
- Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)

Data must be accompanied by sufficient usage metadata to enable their reliable reuse. Some of this information (such as spatial-temporal co-ordinates, parameter labels and units of measure) may be embedded within the data files. The remainder should be included as standard XML documents (e.g. SensorML or ISO19156) or descriptive documents formatted in HTML or PDF.

NERC Data Centre Responsibilities

When a dataset is assigned a DOI, the data centre confirms that:

- The dataset will be made available indefinitely. Note that this does not mean the dataset has to be instantly accessible by requesters (it may have to be retrieved from long-term archiving on tape for example) but that it does have to hold the same data as when its DOI was first minted
- There will be bit-wise fixity of the dataset
- There will be no additions or deletions of files or records
- There will be no changes to the directory structure in the dataset 'bundle'
- Upgrades to versions of data formats will result in new editions of datasets.
- The data centre will provide a full catalogue page (the landing or splash page) which will appear when any user clicks on the DOI hyperlink

Landing Pages

DataCite's key requirement for DOIs is that they must resolve to an 'open access' landing page that describes the dataset.

It would be inappropriate to have the DOI resolve at the archive level ie, giving direct access to the files in the dataset because:

- Users would land with only the information about the dataset, that is just a list of filenames which makes it difficult to be certain of the correct identity of the set
- If the archive structure is changed, it requires re-mapping of all the DOIs

Users are used to seeing landing pages when selecting DOIs since it is common practice with online journals.

DOI landing pages will be the first experience many users will have of a data centre's metadata catalogue and archive. Landing pages should therefore be as user-friendly and easy to understand as possible, and should offer enough human-readable information for users who arrive via a DOI to:

- be confident that they are in the right place to find the dataset they want
- find the data files they want to download

- discover if there are any special requirements of licensing that apply before download
- discover any updates to the cited dataset located elsewhere
- find out any information about the dataset, either by reading an abstract or finding links to other documentation
- know who the author of the dataset is, and whom to credit
- know how to cite the dataset in other publications

Landing pages may also provide machine-readable information about the dataset in other formats such as XML/RDF.

Data centres can change the layout, or add/remove information to/from the landing page at any time, but the user *must* always be able to get to the dataset from the landing page. Certain parts of the metadata describing the dataset should not change, specifically the DOI-mandatory metadata (see next section), as they describe the dataset and represent it faithfully.

If there is a new version of the dataset, a new DOI is needed. The original landing page can indicate that a newer version of the dataset exists (and provide a link to the new version), but the landing page should still point to the original version of the dataset.

Landing pages can have query-based links to other things, for example, papers which cite this dataset, etc.

Structure for Citation and DOI-specific Metadata

The citation and DOI-specific metadata used by the NERC data centres follow the rules laid out in the DataCite metadata schema [12] (currently version 2.1). The schema consists of five mandatory and 12 optional properties which may be used by computers or assembled to create a human-readable citation string. DataCite recommends a particular citation format using the mandatory properties of the metadata scheme:

- Creator (PublicationYear): Title. Publisher. Identifier

DataCite also recommend the following form when information about Version and ResourceType is required:

- Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

This DOI-specific metadata should be automatically generated from the metadata record for the dataset.

Versioning and Granularity of Datasets

The fundamental principle held by the NERC data centres when assigning DOIs to datasets is that datasets are stored unchanged for an indefinite period. Should changes

to a dataset be required, then this will be implemented by publishing a new version of the dataset, which will involve the following:

- Assignment of a new version number (a simple integer sequence that does not support the concepts of major and minor upgrades)
- Assignment of a new DOI
- Creation of a landing page for the new version of the dataset that includes its full version history
- Modification of the landing page of the previous version of the dataset to provide a link to the new version
- Storage of the new dataset in addition to previous versions

Caution needs to be exercised in assigning DOIs to updated versions of datasets as the overhead involved in storing multiple versions of the same dataset which differ only slightly (but have been assigned different DOIs) will rapidly become prohibitive.

It is not necessary (and in fact should be avoided) to assign DOIs to every measurement taken in a dataset. Common sense should apply as to how 'thinly sliced' a dataset should be – we want to avoid the concept of 'minimum publishable unit' being applied to datasets! DOIs should be assigned to datasets which are scientifically meaningful; hence the size of these datasets will vary, according to the scientific domain of the data.

Citation versus Referencing

It is of course possible to cite smaller chunks of the dataset, while using the DOI attached to the complete dataset. For example:

> Bloggs, Jane and Doe, John, Years 2001, 2005 and 2009 from "Our really important measurements of birds in our garden, 2000-2010" doi:10.12345/abcdefg.

It might be helpful to think of the book/chapter/verse analogy. DOIs provide citation at the level of the book, but further information allows the user/reader to get to exactly the required verse. If the dataset is properly frozen, then the reference to a part of it will be easy to find and extract.

The NERC data centres draw a distinction between:

> Citation – where there is a data centre commitment regarding fixity, stability, permanence etc. of a dataset, which is demonstrated by DOI assignment.

and

> Referencing – where there is no data centre commitment regarding fixity, stability, permanence etc. of a dataset. The dataset can still be referenced and found via URL – but the link might be broken and the data may have changed since the reference was written.

Citing Changing Datasets

As mentioned earlier, it is possible to reference an unfrozen dataset, i.e. one that is still being updated, through the use of a citation string with a URL. There is no guarantee to the user of the citation that the dataset retrieved on one date will be the same as it was when the citation was written at some earlier point in time.

The NERC data centres recognise that there are many datasets which become scientifically significant before the dataset is completed and frozen. For this reason, we have come up with the following guidelines for dynamic datasets.

For datasets which are long term and are updated solely by appending new data/files to previous records (e.g. instruments which have been in place for years) and where it is anticipated they will be collecting data for the foreseeable future, it is possible to break the dataset into smaller parts and assign DOIs to those parts. When the dataset is completed in the future, a single DOI can then be issued for the whole dataset. For example, a long-term rain gauge time series spanning 10 years can have a DOI assigned to each year as the year is completed, and then can have a final DOI assigned to the entire time series once the instrument is moved to a new location or taken out of service.

For those datasets subject to both continual data updates as well as data additions, it would be appropriate to take a fixed snapshot of the dataset and store it elsewhere in the repository, and then assign the DOI to that particular snapshot. The snapshot would then be the stable dataset to which the DOI refers. The frequency of these snapshots would be determined by the hosting data centre, depending on such factors as the size of the dataset and its update frequency. A balance would need to be struck between the costs associated with storing multiple snapshots of the same dataset, versus the convenience for the citer in being able to cite exactly the data used.

Next Steps: Data Publication

The NERC Data Citation and Publication Project has primarily been focusing on data citation, as it is something where the mechanisms can be (relatively) easily set up in-house. Dataset Publication (and associated scientific peer review) could also be done in the same way, though it is outside the core remit of the data centres, and therefore it makes far more sense to team up with academic publishers in order to take advantage of the systems they already have in place.

For this reason, we are in consultation with recognized academic publishers to pilot and promote data journals. This work has developed to the extent that a new data journal, *Geoscience Data Journal* (GDJ) will be launched in 2012 in partnership between the Royal Meteorological Society and Wiley-Blackwell. GDJ will join pre-existing data journals, including *Earth System Science Data* [13] and *Geochemistry, Geophysics, Geosystems* (G3) [14].

Publication of datasets, with its associated need for scientific peer review, will bring up other challenges, mainly along the lines of how one would peer review a dataset in the first place. The CLADDIER and OJIMS projects attempted to address this problem and their conclusions will provide an excellent starting point for implementing data peer review and Publication in the future.

Conclusion

The NERC Data Citation and Publication Project has been running since April 2010. At time of writing, 15 datasets in the NERC data centres have been issued with DOIs. We can therefore cite our datasets, giving academic credit to those scientists who are cited – making it more likely they would give us good-quality data to archive, and thereby improving transparency and traceability of the scientific record.

Phase 2 of the project began in November 2011 and will last two years. At the end of this phase, all the NERC data centres will have:

- At least one dataset with associated DOI
- 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 data centres will cite.

Our users are already expressing an interest in data citation; this is an idea whose time has come!

References

1. Bryan Lawrence, Catherine Jones, Brian Matthews, Sam Pepler, Sarah Callaghan. "Citation and Peer Review of Data: Moving Towards Formal Data Publication". *International Journal of Digital Curation*, Vol 6, No 2 (2011) http://www.ijdc.net/index.php/ijdc/article/view/181
2. Sarah Callaghan, Sam Pepler, Fiona Hewer, Paul Hardaker, Alan Gadian. "How to Publish Data Using Overlay Journals: The OJIMS Project". October 2009, Ariadne Issue 61 http://www.ariadne.ac.uk/issue61/callaghan-et-al/
3. Sarah Callaghan, Fiona Hewer, Sam Pepler, Paul Hardaker and Alan Gadian. "Overlay Journals and Data Publishing in the Meteorological Sciences". July 2009, *Ariadne* Issue 60 http://www.ariadne.ac.uk/issue60/callaghan-et-al/
4. Details of Event SCOR/IODE Workshop on Data Publishing, 17 - 18 June 2008, Oostende, Belgium http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=273
5. Details of Event Second SCOR/IODE Workshop on Data Publishing, 9 - 11 March 2009, Oostende, Belgium http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=435
6. Details of Event SCOR/IODE/MBLWHOI Library Workshop on Data Publication, 2 April 2010, Paris, France http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=625
7. Details of Document IOC Workshop Report No. 244, SCOR/IODE/MBLWHOI Library Workshop on Data Publications, 4th Session http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=8098
8. It should be noted that, as yet, there is not general agreement on how to peer-review a set, though discussions are being carried out between interested parties in

an effort to establish a clear and internationally agreed approach to do this. This is discussed later in this article.

9. It is not clear at this time how exactly a 'trusted' data repository would be defined. Several initiatives such as the Data Seal of Approval and World Data System data centre accreditation are in their early stages and could provide guidance on this. In the meantime, the reputation of the data centre in its field provides an indication of how trusted it is, and therefore how stable and secure data hosted by them is considered.
10. Data Publisher for Earth & Environmental Science http://pangaea.de/
11. Earth System Science Data: The Data Publishing Journal http://www.earth-system-science-data.net/
12. DataCite Metadata Schema Repository http://schema.datacite.org
13. *Earth System Science Data* http://earth-system-science-data.net/
14. *Geochemistry, Geophysics, Geosystems* http://www.agu.org/journals/gc/

Author Details

Sarah Callaghan
Senior Scientific Researcher and Project Manager
British Atmospheric Data Centre

Email: sarah.callaghan@stfc.ac.uk
Web site: http://badc.nerc.ac.uk

Sarah Callaghan is project manager for the NERC SIS Data Citation and Publication Project. She has both created and managed datasets in the past and has a keen appreciation of the amount of effort required in both roles!

Roy Lowry
Technical Director
British Oceanographic Data Centre

Email: rkl@bodc.ac.uk
Web site: http://www.bodc.ac.uk

Roy Lowry has an extensive background in the technical aspects of data management and has been involved in many national and international projects including the NERC DataGrid (part of the e-Science Programme) SeaDataNet, and Marine Metadata Interoperability (MMI).

David Walton
Emeritus Professor
British Antarctic Survey

Email: dwhw@bas.ac.uk
Web site: http://www.antarctica.ac.uk/

Professor David Walton was Head of the Environment and Information Division of the British Antarctic Survey before retiring. He is Editor-in-Chief of the scientific journal *Antarctic Science* and has contributed to, compiled and edited six books on research in Antarctica and elsewhere.

_____Article ends_____

Section 3: Keywords

Data citation, data publication, data centres