UNESCO

United Nations
Educational, Scientific and
Cultural Organization

Intergovernmental
Oceanographic
Commission

Ocean Data Publication Cookbook

UNESCO

OCEAN DATA STANDARDS

jcomm IODE

Ocean Data Publication Cookbook

IOC Manuals and Guides 64
Version 1
March 2013

EXECUTIVE SUMMARY

This "Cookbook" has been written for data managers and librarians who are interested in assigning a permanent identifier to a dataset for the purposes of publishing that dataset online and for the citation of that dataset within the scientific literature. A formal publishing process adds value to the dataset for the data originators as well as for future users of the data. Value may be added by providing an indication of the scientific quality and importance of the dataset (as measured through a process of peer review), and by ensuring that the dataset is complete, frozen and has enough supporting metadata and other information to allow it to be used by others. Publishing a dataset also implies a commitment to persistence of the data and allows data producers to obtain academic credit for their work in creating the datasets. One form of persistent identifier is the Digital Object Identifier (DOI). A DOI is a character string (a "digital identifier") used to provide a unique identity of an object such as an electronic document. Metadata about the object is stored in association with the DOI name and this metadata may include a location where the object can be found. The DOI for a document is permanent, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the metadata for the DOI to link to the new URL. A DOI may be obtained for a variety of objects, including documents, data files and images. The assignment of DOIs to peer-reviewed journal articles has become commonplace. This cookbook provides a step-by-step guide to the data publication process and showcases some best practices for data publication. This cookbook is an outcome of the 5th session of the SCOR/IODE/MBLWHOI Library Workshop on Data Publication.

TABLE OF CONTENTS

ANNEXES

Data publication "best practice" examples

1- MBL-WHOI

2- BODC Published Data Library (including example DataCite XML record)

1. DATA PUBLICATION

1.1 What is "Data Publication"?

It is possible to "publish" data relatively easily: at its most basic all a researcher has to do is to put the files on a website somewhere. This makes the data accessible, but without any form of long-term commitment. There are no guarantees that the data will still be in the same location in six months, or that the files haven't become 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. Even if the dataset is readable and has sufficient metadata, there is no information about the scientific quality of the dataset, other than that associated with the creator's reputation.

By contrast, a formal "publishing" process adds value to the dataset for the data originator and for future users of the data. Publishing may provide an indication of the scientific quality and importance of the dataset (as measured through a process of peer review), and 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" also implies a commitment to persistence of the data and allows data producers to obtain academic credit for their work in creating the datasets.

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 Climategate scandal showed that the general public has an interest in the scientific work that government money is funding. Indeed, in the United Kingdom the government wishes to make all data from publicly funded research available to the public for free[1], as does the U.S. National Science Foundation.

1.2 Why Data Publication?

Previously, there was little benefit to a scientist in making their dataset available as a free download from a webpage, unless they worked in certain areas of science where this is expected (e.g., for genetic sequence data from GenBank). In fact, prior to this, the reputational risk of doing so (others might find errors, or 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. However, if the dataset author could receive full citation credit for their data collection effort, thus contributing to measurable performance metrics, motivation for data publication would be greatly increased. Additionally, funding agencies are requiring that data gathered through funded research be made accessible. These funding agencies are requesting a Data Management Plan to be submitted in grant proposals. Thus, data centres are working with scientists to bring data from the closed servers and CDs into archives where they can be properly curated, with the eventual aim of publication and the dataset author receiving full academic credit for their efforts.

The advent of funding agency mandates for open data, such as the National Science Foundation requirement that a data management plan be included in proposals and

---

[1]   http://www.guardian.co.uk/science/2012/jul/15/free-access-british-scientific-research

the European Commission's recent recommendation for open access to scientific data, is expected to provide incentive for authors to make data available. The Scholarly Publishing & Academic Resources Coalition has recently published guidelines on implementing an Open Data Policy []. The ability to cite one's data accurately makes openness advantageous to research scientists.

The assignment of persistent identifiers, specifically Digital Object Identifiers (DOIs), enables accurate data citation. Data publication that enables data citation can certainly be an incentive to make data accessible.

Marine science librarians are increasingly becoming an integral partner in this data publication process. Through their knowledge of cataloguing and metadata creation, and their role in maintaining institutional repositories, they play an important part in making data accessible to all those interested in scientific information and data: science colleagues, policy makers, academics and the interested public. The professional librarian has long-standing knowledge of the publishing process, the new online scholarly information cycle, and the standards necessary to link data to publications.

2. TECHNOLOGY REQUIREMENTS

In this section, we identify the necessary technology to build an operational system for data publication. The annex to this document contains descriptions of some example operational systems utilizing these technologies.

2.1 DSpace

The MBLWHOI selected DSpace as the platform for an Institutional Repository (IR) in 2005. The decision was partially based on the fact that, at the time, it was one of the few IR platforms that would accept a variety of file types.   There were several years when support for DSpace seemed to be waning, but the merger with Duraspace has led to increased support and more widespread use of the product. Because expertise was not available in Woods Hole to support the necessary Java programming to customize and update DSpace, support of DSpace was outsourced to a vendor, Longsight. This approach has been an economical and effective method of support. In 2012, the vendor @mire, which has a history of working with the DuraSpace community, was hired to develop code that enables item-level versioning. @mire submitted the code for review to be included in version DSpace 3.0. The code for item-level versioning was accepted and is included in the most recent release of DSpace.

2.2 Customizing DSpace

Installations of DSpace may be customized to allow additional metadata schema to be incorporated. There are detailed instructions concerning this in the MBLWHOI use case in the Annex.

2.3 File storage options

For those who cannot, or do not, wish to install DSpace on their systems there are other file storage options available. These include File Transfer Protocol (FTP) servers and storage areas which are accessible to web servers and which may deliver files through the Hyper Text Transfer Protocol (HTTP).

File Transfer Protocol (FTP) is a standard network protocol used to transfer files from one host to another host over a Transmission Control Protocol (TCP)-based network, such as the Internet. FTP is built on a client-server architecture and uses separate control and data connections between the client and the server. FTP users may authenticate themselves using a clear-text sign-in protocol, normally in the form of a username and password, but can connect anonymously if the server is configured to allow it. For secure transmission that hides (encrypts) the username and password, and encrypts the content, FTP is often secured with SSL/TLS ("FTPS"). SSH File Transfer Protocol (SFTP) is sometimes used instead, but is technologically different.[2] Most web browsers can access files on FTP sites, but cannot handle extensions to the basic FTP specification. FTP was not designed with high security in mind, and indeed should not be used for files that must be kept secure. Open-source implementations of FTP server software exist for most operating systems and include FileZilla Server; War FTP Daemon; ProFTPD; Pure-FTPd; vsftpd; and wuftpd.

The Apache web server can serve files from within its own directory structure, or may be configured to serve files from a remote server using Apache's rewrite engine. A configuration file example to allow this latter option is shown below.

```
<VirtualHost *>
    # set server name
    ProxyPreserveHost On
    ServerName www.myserver.com
    # configure static file serving
    DocumentRoot /remoteserver/appname/web
    <Directory /remoteserver/myserver/web>
        Order deny,allow
        Allow from all
    </Directory>
    # rewrite incoming requests
    RewriteEngine On
    RewriteCond /remoteserver/appname/web%{REQUEST_FILENAME} !-f
    RewriteRule     ^/(.*)$     http://localhost:8080/appname/$1
[proxy,last]
</VirtualHost>
```

For a variety of reasons, the BODC's Published Data Library follows this technical route. The DOIs issued by BODC resolve to static HTML pages served by an Apache web server. This is an interim solution as it requires hand coding each landing page, meaning that every time a new DOI is requested a new page must be written. At the time of writing, a relational database model has been designed for storing the relevant DOI metadata, but it remains unpopulated. Once populated, a web application will be coded to dynamically produce the landing pages from the content of the database, which is much more sustainable in the long term. Data files are referenced from the landing pages, and are stored on a portion of the local area network which is visible to the Apache web server. This allows the files to be accessed easily by end users, but minimizes security concerns. BODC has a high-bandwidth (100Mb per second) connection to the UK's research and education network (Janet), so there is no issue with users downloading data files, including videos, which have been assigned a DOI.

A final option, for libraries or data centres that do not have the web-serving capacity or the bandwidth to produce an operational system of this type, is to use the IODE's Published Ocean Data (POD) site as a host for data files and the web pages to which a DOI resolves. The use of is detailed in the Annex.

---

[2] http://en.wikipedia.org/wiki/File_Transfer_Protocol

3. THE DATA PUBLICATION PROCESS

3.1 What is a persistent identifier?

The persistent identification of digital resources can play a vital role in enabling their accessibility and re-usability over time. However, progress in defining the nature and functional requirements for identifier systems is hindered by a lack of agreement on what identifiers should actually do. To some, an identifier system is strictly a means of providing a unique name to a digital or analogue resource — either globally or locally. To others, identifier systems must also incorporate associated services such as resolution and metadata binding. Specific requirements will differ, but it is vital that institutions seeking to assign permanent identifiers to datasets recognise that the application and maintenance of identifiers forms just one part of an overall digital preservation strategy and responsibility. Without adequate institutional commitment and clearly defined roles and responsibilities, identifiers cannot offer any guarantees of persistence, location, or availability in the long or short terms.

One form of persistent identifier is the Digital Object Identifier (DOI). A DOI is a character string (a "digital identifier") used to uniquely identify an object such as an electronic document. Metadata about the object is stored in association with the DOI name and this metadata may include a location, such as a URL, where the object can be found. The DOI for a document is permanent, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the metadata for the DOI to link to the new URL.

In the academic publishing field, DOIs are assigned to individual articles in order to provide a unique reference to the source of that article online. Here, we present a workflow which allows an analogous assignment of a DOI to the files which comprise a dataset.

Workflow
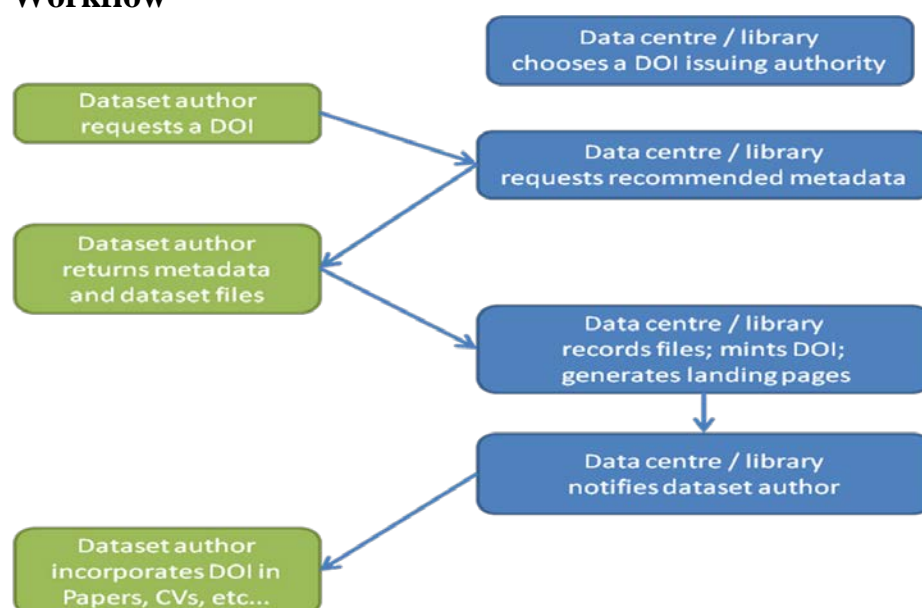
3.2 Choose a DOI-issuing authority

A DOI may be obtained for a variety of objects including documents, data files and images. The first step in obtaining a DOI is to choose a registration organization. The cost will depend on the organization (but in general they are not very expensive). These organizations may also offer additional services, which could add to the cost. One place to start with this process is to contact DataCite who can put you in touch with a local DOI issuing authority. Another route is to contact your institutional library which may already be able to issue DOIs. A list of registration organizations may be found at http://www.doi.org/registration_agencies.html.

The authors have experience in using both the DataCite and CrossRef registration organizations, and the steps involved in obtaining, or "minting", a DOI from each of these bodies is described below.

3.3 Minting a DOI through DataCite

This section is solely for those who hold an account with DataCite and wish to follow a step-by-step guide to minting a DOI. DataCite-issued DOIs may be minted either through a web interface or through an Application Programming Interface (API) if scripts and permissions are set up for the latter. The following steps are aimed at users of the web interface, which allows the minting of one DOI at a time.

1. Navigate to http://mds.datacite.org

2. Log in to the secure part of the website with your username and password

3. Once logged in, select "Register new Dataset"

4. On the resulting page, enter the DOI to be minted in the DOI box. This will take the form:
   a. {your_prefix}/{your_GUID} (GUIDs are explained in more detail below)
   b. e.g. 10.XYZXY/78114093-E2BD-4601-8AE5-3551E62AEF2B

5. Enter the URL of the landing page to which the DOI should point

6. Click "Save" and the DOI will now be minted.

7. Upload an XML file containing the required DataCite metadata. An example can be found in the Annex. The DOI minting process requires this XML file to be available at the time of DOI minting.

Metadata
In addition to the DataCite Metadata Store XML metadata record (see above, and the Annex), it is also recommended that the DOI landing page comprises a comprehensive metadata record describing the dataset. We explore this in more detail in the "Metadata" section below.

3.4 CrossRef DOI Registration

The MBLWHOI Library has a subscription with CrossRef. The annual cost is US$275 and $0.06 for each dataset DOI deposited.
Upon entry into WHOAS, DOIs are deposited with CrossRef for all appropriate datasets. For the MBLWHOI Library the DOI prefix is 10.1575 and the handle prefix is 1912, such that

- when expressing a DOI in print, use: "doi:10.1575/1912/#" where # equals item number

- when expressing a DOI in a metadata field (e.g. WHOAS, dc.identifier.doi), use: "10.1575/1912/#" where # equals item number

- when expressing a DOI in a MARC record (024), use: "10.1575/1912/#" where # equals item number.

- when linking from a MARC record (856), use http://hdl.handle.net/1912/# (where # is the ID of the record in WHOAS)

- when linking to a record in WHOAS from anywhere, use http://hdl.handle.net/1912/#, or http://dx.doi.org/10.1575/1912/# (where # is the ID of the record in WHOAS)

- when depositing a DOI with CrossRef (resource field), use: "http://hdl.handle.net/1912/# where # equals item number.

It should be noted that:

- DOIs are not deposited for articles (pre-prints, drafts or publisher's version).
- Upon deposit, CrossRef will send a confirming message (success or failure). Deposits that fail to load successfully should be corrected and re-deposited.

XML Generation for CrossRef
Notes:

- confirm authors list in correct order; edit as appropriate.

- confirm <publisher_name> and <publisher_place> reflect correct information; edit as appropriate.

- when registering a DOI for an item with a corporate author, for example, for a selected WHOI technology report, edit the xml file as follows:

```
<contributors>
    <organization sequence="first">
        <contributor role="author">
            Woods Hole Oceanographic Institution
        </contributor>
    </organization>
</contributors>
```

XML test parsing (validation) may be done in a web browser: http://www.crossref.org/06members/55InstructionsforNewSchema.html for users with a CrossRef account. Valid XML files may then be uploaded through a separate web page for DOI registration (http://doi.crossref.org). Note that cookies must be enabled in the browser for this to work.

- Select Upload submissions
- Enter file name, e.g. 1912_3078.xml (Area: Live; Type: Metadata)
- Click Upload

3.5 Creating a GUID

DOIs all follow the same format; a prefix (for example the UK's Natural Environment Research Council's DOI prefix, assigned through DataCite, is 10.5285) followed by a unique string of the DOI minter's choice.

The recommended suffix is a Globally Unique Identifier (GUID) as this is almost guaranteed to be a 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 disadvantages of GUIDs are that they do not look attractive, and there is no data centre branding in the string. Their advantages are 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.

There are GUID generators available for a range of programming languages, for instance the sys_guid() command in Oracle; Java's java.util.UUID.randomUUID() method; UUIDTools in Ruby; and UUID module for Python. Oracle generates GUIDs in sequence, but extensive tests undertaken by BODC have shown that this does not affect their unique character.

The web site <http://www.guidgenerator.com/> provides an alternative GUID generation method for those who require a simpler interface.

3.6 Metadata

In addition to the DataCite Metadata Store or CrossRef XML metadata record (see above, and Annex), it is also recommended that the DOI landing page comprises a comprehensive metadata record describing the dataset. The recommended metadata fields are listed in the table below:

Dataset title **Dublin Core: Title**	A title giving an overview of the dataset

Dataset creators
    Dublin Core: Creator / Author;
 	
        Contributor
	The "authors" of the dataset

Dataset subject **Dublin Core: Subject**	ISO19115 topic category(ies) for the dataset

Dataset abstract
    Dublin Core: Description /
        Description.Abstract
	A descriptive abstract outlining the dataset

Dataset description **Dublin Core: Description**	May be used to provide further details of the dataset

Dataset period
    Dublin Core: Period /
        coverage.temporal
	The time span of the dataset

Dataset spatial coverage
    Dublin Core: coverage /
        coverage.Spatial
	The spatial area a dataset covers. Ideally, a controlled vocabulary such as the SeaVoX Sea Areas Gazetteer should be used to populate this field.

Dataset file format **Dublin Core: format**	The predominant data file format for the dataset – see below for guidelines
Dataset file size(s) **Dublin Core: extent.bytes**	The file size, in bytes, of the data files which make up the dataset being published

Dataset language Dublin Core: Language	The human language the metadata and the dataset are written in (e.g. English)
Dataset discovery metadata record	A link to a standard discovery metadata record describing the dataset (e.g. EDMED, GCMD)
Dataset publisher	The data centre responsible for providing the DOI and publishing the dataset
Dataset publication date Dublin Core: Date.Issued	The date on which the dataset was published with a DOI
Dataset DOI Dublin Core: Identifier	The DOI string which has been assigned to the dataset
Dataset citation text Dublin Core: Bibliographic Citation	The recommended citation text for the dataset (see below for guidelines)
Links to data files & usage metadata	Links to the data files themselves and to documentation describing how to use the data

Data must also be accompanied by sufficient usage metadata to enable its reliable reuse. Some of this (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 – Observations and Measurements) or descriptive documents formatted in HTML or PDF.

3.7 Data file formats

In order that the data files being referenced may persist as long as the identifier which has been assigned, there are some considerations concerning the digital format in which the files are stored. As a general rule, data files which make up a publication dataset must:

- Be stored in a well-documented format that conforms with 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.

- Be stored in a format readable by tools that are freely available now and are likely to remain freely available indefinitely.

- Named in a clear and consistent manner throughout the dataset with filenames (rather than pathnames) that reflect the contents and 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.

- Have parameters in data files labelled either 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 plaintext descriptions.

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

Frameworks, such as [7], exist in which to evaluate the suitability of specific data file formats which may be of concern during the publication process.

3.8 Providing a reference to a published dataset

A recommendation of the text used to provide a citation to the newly published dataset should be made.

This human readable citation string should follow the guidelines laid out in section 2.2 of the DataCite metadata schema (http://schema.datacite.org/).

Because users of this schema are members of a variety of academic disciplines, DataCite remains discipline‑ agnostic concerning matters pertaining to academic style sheet requirements. Therefore, DataCite recommends, rather than requires, a particular citation format. In keeping with this approach, the following is the recommended format for rendering a DataCite citation for human readers using the first five properties of the schema:

> Creator (PublicationYear): Title. Publisher. Identifier

It may also be desirable to include information from two optional properties, Version and ResourceType (as appropriate). If so, the recommended form is as follows:

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

For citation purposes, the Identifier may optionally appear both in its original format and in a linkable, http format, as it is practiced by the Organisation for Economic Co‑ operation and Development (OECD), as shown below.

Regarding the PublicationYear, DataCite recommends, for resources that do not have a standard publication year value, to submit the date that would be preferred from a citation perspective.

Here are several examples:

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo.doi:10.1594/PANGAEA.726855.
  http://dx.doi.org/10.1594/PANGAEA.726855

- Geofon operator (2009): GEFON event gfz2009kciu (NW Balkan Region). GeoForschungsZentrum Potsdam (GFZ).
  doi:10.1594/GFG.GEOFON.gfz2009kciu.
  http://dx.doi.org/10.1594/GFZ.GEOFON.gfz2009kciu

- Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. doi: 10.1594/WDCC/dphase mpeps.
  http://dx.doi.org/10.1594/WDCC/dphase_mpeps

4. USING DATA CITATIONS

Most peer-reviewed journals used by ocean scientists encourage the submission of data and other information to supplement published papers. Presumably, all would allow data to be submitted in the form of a DOI or other persistent identifier.

Data Journals
Some journals, such as "data journals" require the submission of data associated with research papers. *Earth System Science Data* (ESSD) is one example. This open-access journal was created in 2009 to provide a location to describe data sets related to the geosciences and is now in its fifth volume. According to the journal Web site, "It is the aim of ESSD to provide the quality assessment for datasets which already reside in permanent repositories." Data must be assigned an identifier before publication and the data and their presentation is part of the review process. Data files must include standard metadata and be available from a certified repository. ESSD is a journal of Copernicus Publications, which is associated with the European Geosciences Union. Another example of a data journal relevant to ocean sciences is the new *Geoscience Data Journal*, published by Wiley. It has not yet published any papers. *Dataset Papers in Geosciences*, published by Hindawi Publishing Corporation is another new data journal.

An important step for data centres or libraries in relation to data journals is to get on to the journals "accepted repositories" list. Each journal gives instructions about how to get on these lists (see appended table). The most general requirement is that the data centre or library has the ability to mint DOIs.

Research Journals
Copernicus Publications also publishes open access, peer-reviewed research journals related to ocean sciences, such as *Biogeosciences*, *Ocean Science*, *Earth System Dynamics*, and *Advances in Geosciences*.. These journals require that any supplementary material be submitted with the manuscript, but do not require the submission of data in the form of a president identifier. An analysis of *Biogeosciences* showed that the percentage of papers with supplemental material (not necessarily data) increased from 8.3% in 2004 to 28.8% in 2012. For *Ocean Science*, there is no clear trend in the percentage of articles having supplemental material since the journal was started in 2005. Unfortunately, virtually all of the supplemental files are in pdf format or in zip files, so they are not machine-readable.

Journals published by the American Geophysical Union (AGU) allow submission of "supporting material", including data files. The instructions to authors requires that any supporting material to be considered with the article is submitted with the article and the publisher (Wiley) will archive and serve the data. Information on submitting supplemental materials can be found at http://publications.agu.org/author-resource-center/author-guide/text-requirements/#supmat. AGU ocean-related journals (*Global Biogeochemical Cycles*; *Geochemistry, Geophysics, Geosystems*; *Geophysical Research Letters*; *Journal of Geophysical Research—Oceans*; and *Paleoceanography*) do not show an upward trend in supplemental information. When considered as a group, the peak in supplemental material (10.4%) occurred in 2005. However, it is not clear how to access supplementary data associated with articles published before the transition to Wiley.

The Association for the Science of Limnology and Oceanography (ASLO) publishes several different journals used by the ocean science community. The best-known of these is *Limnology and Oceanography* (*L&O*). Since 2003, *L&O* has been publishing electronic appendixes that are similar to the supplemental material sections allowed

by other journals. The *L&O* instructions to authors make it clear that the appendixes are only for data tables that would make the paper more understandable, not for raw data. Not all the files contain data, but the ones that do are in Excel or csv format. Articles published in the last three years are locked (unless an extra fee was paid to offer them for free), but the appendix files are unlocked even if the articles are locked. For *L&O: Methods*, the percentage of articles with supplementary material has 4.5 and 12% over the past 10 years.

The Ecological Society of America (ESA) provides a Web location for storage and retrieval of data sets related to articles published in ESA journals (see http://data.esa.org). Fifty data sets are accessible through this site, some of which are related to marine environments.

The following table provides specific information about data journals that might be of interest to ocean scientists.

Name of Data Journal (website)	Aims and Scope	Repository Criteria	Other notes
Geoscience Data Journal www.geoscience data.com	*From http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060/homepage/ProductInformation.html* Geoscience Data Journal provides an Open Access platform where scientific data can be formally published, in a way that includes scientific peer-review. Thus the dataset creator attains full credit for their efforts, while also improving the scientific record, providing version control for the community and allowing major datasets to be fully described, cited and discovered. An online-only journal, GDJ publishes short data papers cross-linked to – and citing – datasets that have been deposited in approved data centres and awarded DOIs. The journal will also accept articles on data services, and articles which support and inform data publishing best practices. Data is at the heart of science and scientific endeavour. The curation of data and the science associated with it is as important as ever in our understanding of the changing earth system and thereby enabling us to make future predictions. Geoscience Data Journal is working with recognised Data Centres across the globe to develop the future strategy for data publication, the recognition of the value of data and the communication and exploitation of data to the wider science and stakeholder communities. Content description: A data article describes a dataset, giving details of its collection, processing, file formats etc., but does not go into detail of any scientific analysis of the dataset or draw conclusions from that data. The data paper should allow the reader to understand the when, why and how the data was collected, and what the data is. Subject coverage: GDJ will accept contributions in the areas of Weather and Climate; Oceanography; Atmospheric and Ocean Chemistry; Cryosphere; Biosphere and Land Surface and Geology. Article publication fee: This is an Open Access journal and there will be a fee to publish – unless this is waived by the Editor for exceptional circumstances.	*http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060/homepage/data_center_faqs.htm* **What would my data center have to do to be approved by GDJ?** The primary requirement is to be able to mint DOIs. Given that data publishing is an evolving field, we are keen to work with fellow stakeholders to promote data publishing and cross-linking (see for instance, the PREPARDE project). Consequently, we expect the process and requirements to develop over time, and to update this FAQ sheet accordingly. **How does my institution go about becoming approved?** If you are able to show you can mint DOIs then the main criteria has been addressed. Other than that, we are looking for evidence that the datasets are being lodged within a long-term sustainable repository, and that it will be possible to put in cross-links so that readers of either the dataset or the Data Paper can move from one site to the other. **Is the repository approval a lengthy process?** No, it should be reasonably straightforward. The key driver to approval is to have a primary contact within the data center with whom we can work. **Is there any support or guidance available from the journal? Whom do I contact to get my repository approved?** Yes. Several members of the Editorial Board have expertise in this area and will be willing to help with specific queries. Please contact the journal at GDJ@wiley.com if you have any questions or would like to get your repository approved. **Once approved what then?** Once your data centre has been approved it will be listed on our approved list here. We also would like to work with you to ensure that all scientists depositing data in your centre are aware of the journal and the facility to submit a data paper. Please therefore provide us with contact details for a named member of staff with whom we can liaise to contact scientists who deposit and discuss link sharing. **Are there any licensing issues I need to be aware of?** Wiley does not claim any rights over datasets residing in repositories, and copyright on any article published by a Wiley Open Access journal is retained by the author(s) under the terms of the Creative	Open access for papers, doesn't mandate open access for datasets.

		Commons Attribution License (CC BY) which allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. Having said this, in the interest of promoting the sharing and re-use of research data, we prefer that reviewers and readers are able to view and re-use the research data with the minimum of restrictions. For more information visit the journal's Open Access License and Copyright page.	
Earth System Science Data http://earth-system-science-data.net/	*From http://earth-system-science-data.net/:* Earth System Science Data (ESSD) is an international, interdisciplinary journal for the publication of articles on original research data(sets), furthering the reuse of high (reference) quality data of benefit to Earth System Sciences. The editors encourage submissions on original data or data collections which are of sufficient quality and potential impact to contribute to these aims. The journal maintains sections for regular length articles, brief communications (e.g., on additions to datasets) and commentary, as well as review articles and "Special Issues". Articles in the data section may pertain to the planning, instrumentation and execution of experiments or collection of data. Any interpretation of data is outside the scope of regular articles. Articles on methods describe nontrivial statistical and other methods employed, e.g. to filter, normalize or convert raw data to primary, published data, as well as nontrivial instrumentation or operational methods. Any comparison to other methods is out of scope of regular articles. Review articles may compare methods or relative merits of datasets, the fitness of individual methods or datasets for specific purposes or how combinations might be used as more complex methods or reference data collections. Earth System Science Data has an innovative two-stage publication process involving the scientific discussion forum Earth System Science Data Discussions (ESSDD), which has been designed to: foster scientific discussion; maximise the effectiveness and transparency of scientific quality assurance; enable rapid publication of new scientific results; make scientific publications freely accessible. In the first stage, papers that pass a rapid access peer-review are immediately published on the Earth System Science Data Discussions (ESSDD) website. They are then subject to Interactive Public Discussion, during	*From http://www.earth-system-science-data.net/general_information/repository_criteria.html* The precondition to submit a manuscript for publication in Earth System Science Data (ESSD) and its scientific discussion forum Earth System Science Data Discussions (ESSDD) is that the data sets referenced in the manuscript are submitted to a long-term repository. The following basic criteria have to be fulfilled under all circumstances: Persistent Identifier: The data sets have to have a digital object identifier, e.g. doi. Open Access: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge. Liberal Copyright: Anyone must be free to copy, distribute, transmit and adapt the data sets as long as he/she is giving credit to the original authors (equivalent to the Creative Commons Attribution License). Long-term Availability: The repository has to meet the highest standards to guarantee a long-term availability of the data sets and a permanent access.	Open access papers and mandates open access for data.

	which the referees' comments (anonymous or attributed), additional short comments by other members of the scientific community (attributed) and the authors' replies are also published in ESSDD. In the second stage, the peer-review process is completed and, if accepted, the final revised papers are published in ESSD. To ensure publication precedence for authors, and to provide a lasting record of scientific discussion, ESSDD and ESSD are both ISSN-registered, permanently archived and fully citable. Earth System Science Data also offers an efficient new way of publishing special issues, in which the individual papers are published as soon as available and linked electronically (for more information see Special Issues).		
Ecological Archives - Data Papers http://esapubs.org /archive/archive_ D.htm	**What is Ecological Archives?** Ecological Archives publishes materials that are supplemental to articles that appear in the ESA journals (Ecology, Ecological Applications, Ecological Monographs, Ecosphere, and Bulletin of the Ecological Society of America), as well as peer-reviewed data papers with abstracts published in the printed journals. Ecological Archives is published in digital, Internet-accessible form. Three kinds of publications appear in Ecological Archives: appendices, supplements, and data papers. Ability to publish appendices and supplements in Ecological Archives allows authors to shorten the paper version of a manuscript without withholding germane material not essential for understanding the paper. It also allows authors to make available substantial amounts of supporting material such as methodological details, data tables, graphs illustrating additional analyses, photographs, additional references, and supplemental discussion, all as citable entities.	**Data Registry** In addition, all authors are encouraged to register their data at ESA's official Data Registry at data.esa.org The Data Registry simply serves to announce the existence of data and to provide contact information. By registering data, one does not relinquish rights to research findings. In fact, the registry may serve to establish precedence for ecological studies. Our hope is the the Data Registry will eventually be linked to data archives containing the actual data referred to in the registry, and that all data underlying published papers in ESA journals will be readily available for purposes of verification, replication, and meta-analysis. The ESA Data Registry form is for registering data sets associated with articles published in the journals of the Ecological Society of America. Other Ecological data sets can be registered with the Knowledge Network for Biocomplexity (KNB).	Page about what ESA considers a data paper and guidelines for reviewers at http://esapubs.org/archive/instruct_d.htm Doesn't seem to give any information about what constitutes an appropriate data repository.
Hindawi publishing: http://www.datase ts.com/ Dataset Papers in Agriculture Dataset Papers in Biology Dataset Papers in Chemistry Dataset Papers in Ecology Dataset Papers in Geosciences Dataset Papers in Materials Science Dataset Papers in Medicine Dataset Papers in Nanotechnology Dataset Papers in	*The following seems to cover all the dataset journals:* Dataset Papers in Geosciences is a peer-reviewed, open access journal devoted to the publication of dataset papers in all areas of geosciences. Dataset Papers in Geosciences is part of a series of journals devoted to the dissemination of dataset papers covering a wide range of academic disciplines. In addition to publishing dataset papers, the journal hosts the underlying data that is associated with these papers and makes it accessible to all researchers worldwide.	None – the journal hosts the underlying dataset	

Neuroscience Dataset Papers in Pharmacology Dataset Papers in Physics			
Biodiversity Data Journal http://www.pensoft.net/journals/bdj/	*From http://www.pensoft.net/journals/bdj/about/Focus%20and%20Scope#Focus and Scope* Biodiversity Data Journal (BDJ) is a community peer-reviewed, open-access, comprehensive online platform, designed to accelerate publishing, dissemination and sharing of biodiversity-related data of any kind. All structural elements of the articles – text, morphological descriptions, occurrences, data tables, etc. – will be treated and stored as DATA, in accordance with the Data Publishing Policies and Guidelines of Pensoft Publishers. The journal will publish papers in biodiversity science containing taxonomic, floristic/faunistic, morphological, genomic, phylogenetic, ecological or environmental data on any taxon of any geological age from any part of the world with no lower or upper limit to manuscript size. For example: single taxon treatments and nomenclatural acts (e.g., new taxa, new taxon names, new synonyms, changes in taxonomic status, re-descriptions, etc.); data papers describing biodiversity-related databases, including ecological and environmental data; any kind of sampling report, local observations or occasional inventories; local or regional checklists and inventories; habitat-based checklists and inventories; ecological and biological observations of species and communities; any kind of identification keys, from conventional dichotomous to multi-access interactive online keys; descriptions of biodiversity-related software tools.	*from http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf* How to Publish Data At Pensoft, data can be published either (a) as supplementary files related to a research paper or (b) in association with a stand-alone description of the data resource (a Data Paper). Within these two main routes, Pensoft supports the following methods for data publishing: • Supplementary data files (max. 20 MB each) related to and published with a research article, and stored on the publisher's website. Guidelines on the format are available here. • Primary biodiversity data (species-by-occurrence records) published through the GBIF Integrated Publishing Toolkit (IPT). A format that is strongly encouraged for the publication of biodiversity and species occurrence data, checklists and their associated metadata is the Darwin Core Archive (DwC-A) format. • Datasets other than primary biodiversity data (e.g., ecological observations, environmental data, genome data and other data types) preserved in certified institutional or international data repositories and linked permanently to a research article or a Data Paper. Best practice recommendations: • Deposition of data in an established international repository is always to be preferred to supplementary files published on a journal's website.8 • Occurrence-by-species records should be deposited through GBIF IPT. • Genomic data should be deposited at GenBank, either directly or via an affiliated repository, e.g. Barcode of Life Data Systems (BOLD). • Phylogenetic data should be deposited at TreeBASE, either directly or through the Dryad Data Repository. • All other biological data, including heterogeneous datasets, should be deposited in the Dryad Data Repository. • Repositories not mentioned above, including institutional repositories, may be used at the discretion of the author. • Digital Object Identifiers (DOIs) or other persistent links (URLs) to the data deposited in repositories, as well as the name of the repository, should always be published in the paper describing that data resource. Other Repositories mentioned • PANGAEA. • The Knowledge Network for Biocomplexity (KNB) • The National Biological Information Infrastructure • DataBasin • DataONE • The PaleoBiology Database • The Research Collaboratory for	Data publishing policies and guidelines give a good overview of what's needed in general for data publishing. http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf

		Structural Bioinformatics (RCSB)'s Protein Data Bank (PDB) • The Universal Protein Resource (UniProt) • INSPIRE.	
Journal of Open Archaeology Data http://openarchaeologydata.metajnl.com/	*From http://openarchaeologydata.metajnl.com/about/* JOAD publishes data papers, which do not contain research results but rather a concise description of a dataset, and where to find it. Papers will only be accepted for datasets that authors agree to make freely available in a public repository. This means that they have been deposited in a data repository under an open licence (such as a Creative Commons Zero licence), and are therefore freely available to anyone with an internet connection, anywhere in the world. A data paper is a publication that is designed to make other researchers aware of data that is of potential use to them. As such it describes the methods used to create the dataset, its structure, its reuse potential, and a link to its location in a repository. It is important to note that a data paper does not replace a research article, but rather complements it. When mentioning the data behind a study, a research paper should reference the data paper for further details. The data paper similarly should contain references to any research papers associated with the dataset. Any kind of archaeological data is acceptable, including for example: geophysical data; quantitative or qualitative data; images; notebooks; excavation data, software, etc.	*From http://openarchaeologydata.metajnl.com/repositories/* The following repositories meet our peer-review requirements and are recommended for the archiving of JOAD datasets. Please contact us if you would like to use another repository or recommend that we add it to our list. <table><tr><th>International repositories</th><th>National repositories</th><th>Institutional repositories</th></tr><tr><td>Archaeology Data Service Figshare Open Context tDAR</td><td>Arachne (Germany) DANS (Netherlands)</td><td>UCL Discovery</td></tr></table> *For each repository, the page also provides extra information on:* *Location* *Focus and suitability* *Cost* *Licenses* *Identifiers used* *Sustainability* *Deposit instructions* *From http://openarchaeologydata.metajnl.com/peer-review/:* 2. The deposited data The repository the data is deposited in must be suitable for this subject and have a sustainability model (see our list of recommended repositories). The data must be deposited under an open license that permits unrestricted access (e.g. CC0, CC-BY). The deposited data must include a version that is in an open, non-proprietary format. The deposited data must have been labelled in such a way that a 3rd party can make sense of it (e.g. sensible column headers, descriptions in a readme text file). The deposited data must be actionable – i.e. if a specific script or software is needed to interpret it, this should also be archived and accessible.	Still in beta Open access for papers, mandates open access for data too. (I like their infographic on benefits of publishing data at http://openarchaeologydata.metajnl.com/about/)
F1000 Research http://f1000research.com	*from http://f1000research.com/about/* Data Publication: F1000 Research promotes publication, refereeing and sharing of full datasets to encourage collaboration and accelerate scientific discovery. Data articles are citable and authors are credited when data are reused.	*Can't find anything, but are partnering with Dryad, biosharing and figshare.* *APCs include up to 1 GB of data. For 1-5 GB of data with an article, an additional US $200 to cover the storage costs is charged. Beyond 5 GB of data, authors are asked to contact F1000R to discuss the costs.*	*The information pack at http://f1000research.com/wordpress/wp-content/uploads/2012/08/F1000R-Online-Information-Pack.pdf gives information about the publication model being used.*

	From http://f1000research.com/wordpress/wp-content/uploads/2012/08/F1000R-Online-Information-Pack.pdf Data Articles: A dataset (or set of datasets) together with the associated methods/protocol used to generate the data. A Data Article may be published as a stand-alone article, or in conjunction with a Research Article		
CODATA's Data Science Journal http://www.codata.org/dsj/index.html	The Data Science Journal is a Journal of the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU) ISSN 1683-1470 The Data Science Journal is a peer-reviewed, open access, electronic journal publishing papers on the management of data and databases in Science and Technology. The scope of the Journal includes descriptions of data systems, their publication on the internet, applications and legal issues. All of the Sciences are covered, including the Physical Sciences, Engineering, the Geosciences and the Biosciences, along with Agricultural and the Medical Sciences. The Journal publishes data or data compilations, if the quality of data is excellent or if significant efforts are required in compilation. The Journal publishes online simulation, database, and other experiments overcoming the inherent limitations of traditional, static print journals, thereby adding an entirely new dimension to the communication and exchange of data research results and educational materials. *Scope of the journal is a long list at* http://www.codata.org/dsj/scope.html	*It looks like data is stored as part of the supplemental information* http://www.codata.org/dsj/submissions.html Supplemental data and information in any format (BMP, JPEG, DOC, XLS, WAV, VOB, etc.) are acceptable. Supporting materials may take the form of figures, tables, datasets, videos, etc. Supporting material is reviewed along with the paper, and is referred to in the text. Authors cannot alter the information after acceptance for publication.	
BMC Research Notes http://www.biomedcentral.com/bmcresnotes/	BMC Research Notes is an open access journal publishing scientifically sound research across all fields of biology and medicine. The journal provides a home for short publications, case series, and incremental updates to previous work with the intention of reducing the loss suffered by the research community when such results remain unpublished. BMC Research Notes also encourages the publication of software tools, databases and data sets and a key objective of the journal is to ensure that associated data files will, wherever possible, be published in standard, reusable formats. We are currently working with researchers across the full spectrum of biomedical research to define appropriate recommendations for domain-specific data file standards.	*From* http://www.biomedcentral.com/bmcresnotes/authors/instructions/datanote Publishing Datasets Through a special arrangement with LabArchives, LLC, authors submitting manuscripts to BMC Research Notes can obtain a complimentary subscription to LabArchives with an allotment of 100MB of storage. LabArchives is an Electronic Laboratory Notebook which will enable scientists to share and publish data files in situ; you can then link your paper to these data. Data files linked to published articles are assigned digital object identifiers (DOIs) and will remain available in perpetuity. Use of LabArchives or similar data publishing services does not replace preexisting data deposition requirements, such as for nucleic acid sequences, protein sequences and atomic coordinates. Instructions on assigning DOIs to datasets,	http://www.biomedcentral.com/bmcresnotes/authors/instructions/datanote Instructions for authors about data notes

| | | so they can be permanently linked to publications, can be found on the LabArchives website. Use of LabArchives' software has no influence on the editorial decision to accept or reject a manuscript.

Authors linking datasets to their publications should include an Availability of supporting data section in their manuscript and cite the dataset in their reference list. | |

5. USE OF DATA CITATION BY INSTITUTIONAL MANAGEMENT OR FUNDING AGENCIES

For the process of data publication to catch on, an important development will be the use of data publications for promotion and tenure decisions in universities and other institutions. The ability to assign DOIs to data publications is only the first step in this process. The next step will be for listing of data publications on CVs to become routine and for indexing services to include data publications in their listings of citations and calculations of "h scores" and other metrics of scientific output.

REFERENCES

[1]   *SCOR/IODE/MBLWHOI Library Workshop on Data Publication, 4th Session, British Oceanographic Data Centre, Liverpool, United Kingdom, 3-4 November 2011*. Paris, UNESCO, 14 November 2011 (IOC Workshop Report No. 244) (English)

[2]   Sarah Callaghan, Roy Lowry, David Walton and members of the Natural Environment Research Council Science Information Strategy Data Citation and Publication Project team. March 2012. *Data Citation and Publication by NERC's Environmental Data Centres*. Ariadne Issue 68. http://www.ariadne.ac.uk/issue68/callaghan-et-al

[3]   *Data Publisher for Earth & Environmental Science* http://pangaea.de/

[4]   *Earth System Science Data: The Data Publishing Journal* http://www.earth-system-science-data.net/

[5]   *DataCite* http://www.datacite.org/

[6]   *Geoscience Data Journal* http://eu.wiley.com/WileyCDA/WileyTitle/productCd-GDJ3.html

[7]   N. Lormant, C. Huc, D. Boucon and C. Miguel (2005) *How to evaluate the ability of a file format to ensure long-term preservation for digital information?* In *Proceedings of PV2005: Ensuring Long-Term Preservation and adding Value to Scientific and Technical Data*, Royal Society of Edinburgh, Edinburgh, UK. 21st-23rd November 2005.

[8]   *Geochemistry, Geophysics, Geosystems* http://www.agu.org/journals/gc/

[9]   Greg Tananbaum (2013) *Implementing an Open Data Policy: A Primer for Research Funders* Scholarly Publishing and Academic Resources Coalition (SPARC). http://www.arl.org/sparc/bm~doc/sparc-open-data-primer-final.pdf

ANNEX : Data publication "best practice" examples

1. MBL-WHOI

Qualified Metadata for Datasets

Metadata elements	Note	Example	In simple item record
dc.contributor.author	Repeat for multiple authors	Johnson, Mark	Authors: Johnson. Mark Doe, J. Q. Other
dc.contributor.author	Repeat for multiple authors	Doe, J. Q.	
dc.contributor.other		Use for "non-author" contributors to datasets	
dc.coverage.spatial		29.7182N 17.9012W	Location: 29.7182N 17.9012W El Hierro, Canary Islands, Spain
dc.coverage.spatial		El Hierro, Canary Islands, Spain	
dc.coverage.temporal		11:34:37 Local	
dc.date.accessioned	Created by WHOAS	2010-05-01	
dc.date.created	Date of creation	2005-10-21	Created: 21-Oct-2005
dc.date.issued	Date of publication	2010-06-01	Date: 06-June-2010
dc.description.abstract		Recording of untagged beaked whales made by a DTAG floating at depth. Location: El Hierro, Canary Islands, Spain. Permit: Canary Islands Government permit to University of La Laguna. Depth: 595 m (start) to 236 m (end). Water depth: approx. 904 m Quality: excellent at start,	Abstract: Recording of untagged beaked whales made by a DTAG floating at depth. Location: El Hierro, Canary Islands, Spain. Permit: Canary Islands Government permit to University of La Laguna. Depth: 595 m (start) to 236 m (end). Water depth: approx.

		poor at end. Species: Mesoplodon densirostris; Blainville's beaked whale. Quantity: 3	904 m. Quality: excellent at start, poor at end. Species: Mesoplodon densirostris; Blainville's beaked whale. Quantity: 3
dc.description		Put process by which data collected here. Sampling rate 192 kHz; Channels: 1; Resolution: 16 bit; Compression: uncompressed; Recording device: DTAG serial number 214; Analog-to-digital converter: CS5341 (sigma-delta). Sensitivity - 171 dB re microPascal per V (clipping level); Filter: 500 Hz 1-pole analog high pass filter	Description: Sampling rate 192 kHz; Channels: 1; Resolution: 16 bit; Compression: uncompressed; Recording device: DTAG serial number 214; Analog-to-digital converter: CS5341 (sigma-delta). Sensitivity - 171 dB re microPascal per V (clipping level); Filter: 500 Hz 1-pole analog high pass filter
dc.format.extent	Repeat for multiple files	nn bytes	
dc.format.mimetype	Repeat for multiple files	txt, jpg, psv, etc	
dc.identifier.citation	Citation of article	*ICES Journal of Marine Science*. 67 (2010): 583-593.	
dc.identifier.other		doi:10.1575/1912/[item#]	
dc.identifier.uri	Created by WHOAS	http://hdl.handle.net/1912/[item #]	URI: http://hdl.handle.net/1912/[item #]
dc.relation.haspartof	Repeat for multiple datasets	http://hdl.handle.net/1912/[item #]	Relation (has part of): http://hdl.handle.net/1912/[item #]
dc.relation.ispartof	Link to Article	http://hdl.handle.net/1912/[item #]	Relation (is part of): http://hdl.handle.net/1912/[item #]
dc.rights.uri		http://creativecommons.org/licenses/by-nc/2.5	Rights: http://creativecommons.org/licenses/by-nc/2.5
dc.source.uri	Link to other data	http:// …..	TBD: http://

		providers ?	
dc.subject	Repeat for multiple terms/phrases	Mesoplodon densirostris	
dc.subject	Repeat for multiple terms/phrases	Blainville's beaked whale	
dc.title		Recording of untagged beaked whales at El Hierro, Canary Islands, Spain	Title: Recording of untagged beaked whales at El Hierro, Canary Islands, Spain
dc.type		Data set	
dwc.scientificName			
dwc.genus			

Additional information coverage:
dc.coverage.temporal
```
>       Year:
>           YYYY (eg 1997)
>       Year and month:
>           YYYY-MM (eg 1997-07)
>       Complete date:
>           YYYY-MM-DD (eg 1997-07-16)
>       Complete date plus hours and minutes:
>           YYYY-MM-DDThh:mmTZD (eg 1997-07-16T19:20+01:00)
>       Complete date plus hours, minutes and seconds:
>           YYYY-MM-DDThh:mm:ssTZD (eg 1997-07-16T19:20:30+01:00)
>       Complete date plus hours, minutes, seconds and a decimal
fraction of a
> second
>           YYYY-MM-DDThh:mm:ss.sTZD (eg 1997-07-16T19:20:30.45+01:00)
>
> where:
>
>       YYYY = four-digit year
>       MM   = two-digit month (01=January, etc.)
>       DD   = two-digit day of month (01 through 31)
>       hh   = two digits of hour (00 through 23) (am/pm NOT allowed)
>       mm   = two digits of minute (00 through 59)
>       ss   = two digits of second (00 through 59)
>       s    = one or more digits representing a decimal fraction of a
second
>       TZD  = time zone designator (Z or +hh:mm or -hh:mm)
```

Dublin Core information for bounding box http://dublincore.org/documents/dcmi-box/
Example from WHOAS:
dc.coverage.spatial westlimit: -67.9328; southlimit: 28.6933; eastlimit: -57.1648; northlimit:
35.8337

Simple item record

Title:
Recording of untagged beaked whales at El Hierro, Canary Islands, Spain
Authors:
Johnson. Mark
Doe, J. Q.
Created:
21-Oct-2005
Location:
29.7182N 17.9012W
Abstract:
Recording of untagged beaked whales made by a DTAG floating at depth. Location: El
Hierro, Canary Islands, Spain. Permit: Canary Islands Government permit to University of La
Laguna. Depth: 595 m (start) to 236 m (end). Water depth: approx. 904 m. Quality: excellent
at start, poor at end. Species: Mesoplodon densirostris; Blainville's beaked whale. Quantity: 3
URI:
http://hdl.handle.net/1912/[item #]
Appears in Collections:
DTAG

Title: md05_294a10590-11850.wav
Creator: Johnson, Mark, Woods Hole Oceanographic Institution
Subject: audio recording - beaked whale
Type: sound
Description: recording of untagged beaked whales made by a DTAG floating at depth
        Location: El Hierro, Canary Islands, Spain
        Permit: Canary Islands Government permit to University of La Laguna
        Depth: 595 m (start) to 236 m (end)
        Water depth: approx. 904 m
        Quality: excellent at start, poor at end
        Species: Mesoplodon densirostris; Blainville's beaked whale
        Quantity: 3;
Date: 2005-10-21; 17:31:01 Local
Format[IMT]: audio/wav
Format[Extent]: nn bytes; 21.0 minutes
Digitization Specifications:
        Sampling rate 192 kHz;
        Channels: 1;
        Resolution: 16 bit;
        Compression: uncompressed;
        Recording device: DTAG serial number 214;
        Analog-to-digital converter: CS5341 (sigma-delta)
        Sensitivity -171 dB re microPascal per V (clipping level);
        Filter: 500 Hz 1-pole analog high pass filter
Resource Identifier: – DOI or URI? talk to Colleen Hurter?
Rights Management: http://creativecommons.org/licenses/by-nc/2.5;
        Contact publisher for attribution
Publisher: The DTAG Project, Woods Hole Oceanographic Institution;

Contact: Mark Johnson, majohnson@whoi.edu
Source: excerpted from: md05_294a chips 7-8, cue 10590 – 11850, channel 1
Coverage[Spatial]: 29.7182N 17.9012W (El Hierro, Spain)
Contributing Institution: Woods Hole Oceanographic Institution

Adding Additional Metadata Fields to Item

After an item has been ingested into the repository, its metadata can be edited; fields can be added, changed, or deleted.

These instructions are for DSpace version 1.5.2 using the Manakin interface. Illustrations are taken from a test installation of DSpace maintained for the MBLWHOI Library.

1) Log in as a DSpace administrator.

2) Retrieve an item.

3) Link to "Edit this item"
4) Link to "Item Metadata"
   a. Name: select field from pull-down menu, e.g. dwc.genus
   b. Value: insert text, e.g. "salpa"
   c. Click "Add new metadata"
   d. Repeat for additional fields
   e. Click "Update"
   f. Click "Return" to return to item record

Adding Additional Metadata Schemas to DSpace

DSpace uses the qualified Dublin Core schema as its default metadata registry. Additional schemas, for example, Darwin Core, can be added to the DSpace installation through the Administrator's user interface.

These instructions are for DSpace version 1.5.2 using the Manakin interface. Illustrations are taken from a test installation of DSpace maintained for the MBLWHOI Library.

1) Log in as a DSpace administrator.

2) Link to the Metadata registry
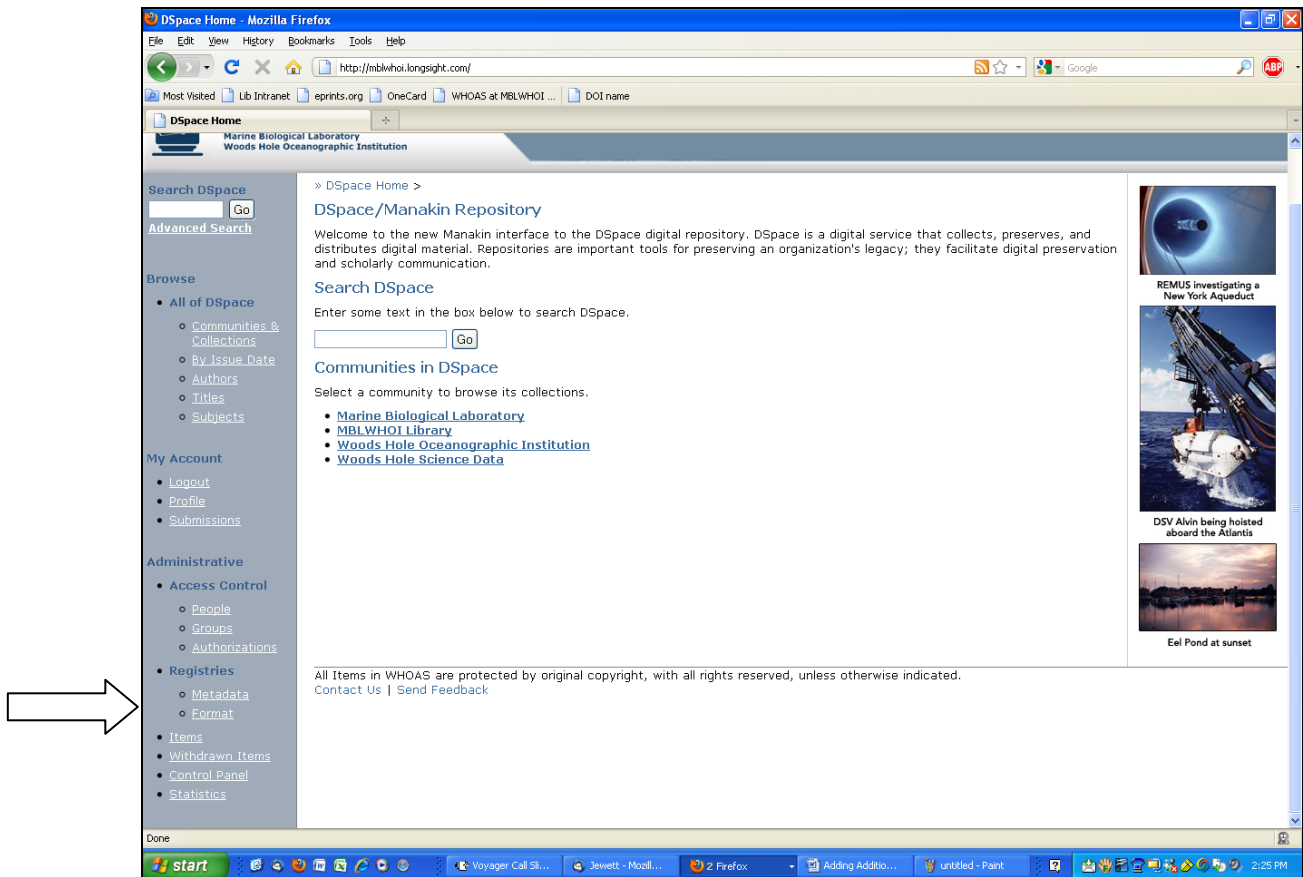
3) "Add a new schema"

a. Namespace: provide the established URI for the new schema, e.g. Darwin Core is found at http://rs.tdwg.org/dwc/terms/

b. Name: provide a shorthand notation for the schema to be used to prefix a field's name, e.g. "dwc" for Darwin Core

c. Click "Add new Schema"
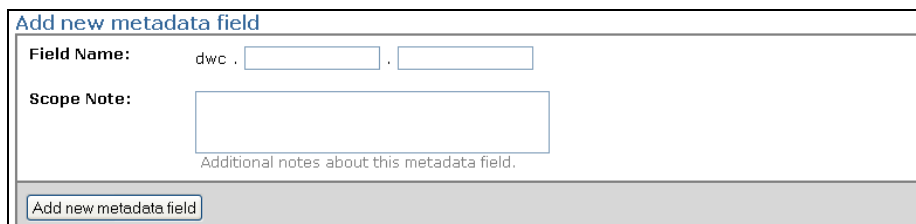
Metadata registry

The metadata registry maintains a list of all metadata fields available in the repository. These fields may be divided amongst multiple schemas. Ho DSpace requires the qualified Dublin Core schema. You may extend the Dublin Core schema with additional fields or add new schemas to the regis
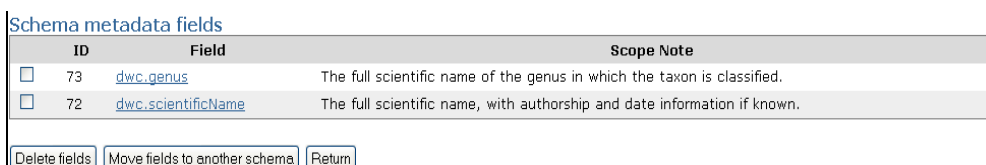
ID	Namespace	Name
1	http://dublincore.org/documents/dcmi-terms/	dc
2	http://rs.tdwg.org/dwc/terms/	dwc

Select the (new) schema to add one or more metadata fields

4) "Add new metadata field"

Add new metadata field

Field Name: dwc . [   ] . [   ]

Scope Note: [   ]

Additional notes about this metadata field.

[Add new metadata field]

    a. Field Name: provide the field name, e.g. "genus"

    b. Scope Note: include notes about the metadata field, e.g. "The full scientific name of the genus in which the taxon is classified."

    c. Click "Add new metadata field"

    d. Repeat for additional metadata fields

Schema metadata fields

	ID	Field	Scope Note
☐	73	dwc.genus	The full scientific name of the genus in which the taxon is classified.
☐	72	dwc.scientificName	The full scientific name, with authorship and date information if known.

[Delete fields] [Move fields to another schema] [Return]

Once the new schema is added, its fields may be used.
Fields may be added to new or existing items in the repository.
Fields may be added to new or existing item (Collection) templates.

Adding Additional Metadata Fields to Item

After an item has been ingested into the repository, its metadata can be edited; fields can be added, changed, or deleted.

These instructions are for DSpace version 1.5.2 using the Manakin interface. Illustrations are taken from a test installation of DSpace maintained for the MBLWHOI Library.

1) Log in as a DSpace administrator.

2) Retrieve an item.

3) Link to "Edit this item"

4) Link to "Item Metadata"
   a. Name: select field from pull-down menu, e.g. dwc.genus
   b. Value: insert text, e.g. "salpa"
   c. Click "Add new metadata"

> d. Repeat for additional fields
>
> e. Click "Update"
>
> f. Click "Return" to return to item record

Providing Access to Embargoed Files

Content in a DSpace repository may be embargoed for period of time. The following procedure outlines the process by which access to embargoed file(s)bitstream(s) by users other than adminstriators is made possible, for example, peer-reviewers needing access to files that support articles submitted for publication.

Users must first register in WHOAS and notify the WHOAS administrator that they have done so. The following information also needs to be provided:

1) Which record(s) do they need access to?

2) For how long?

3) Are they to remain anonymous from the creator/author of the content, that is, are they a reviewer?

Notes:

Once the time period needed to access embargoed item(s) passes, remove the e-persons added to the group and restore the original permissions for the file(s)/ bitstream(s).

If access to multiple embargoed items is needed during the same or overlapping time periods by different groups of people, create multiple groups, such as "read embargo Bell Center", "read embargo BCO-DMO".

Create a new group, or edit an existing group

1) Under Access Control, select Groups

2) Select "Create a new group"

3) Name it , i.e. "read embargo"

4) Add member(s) [E-people]
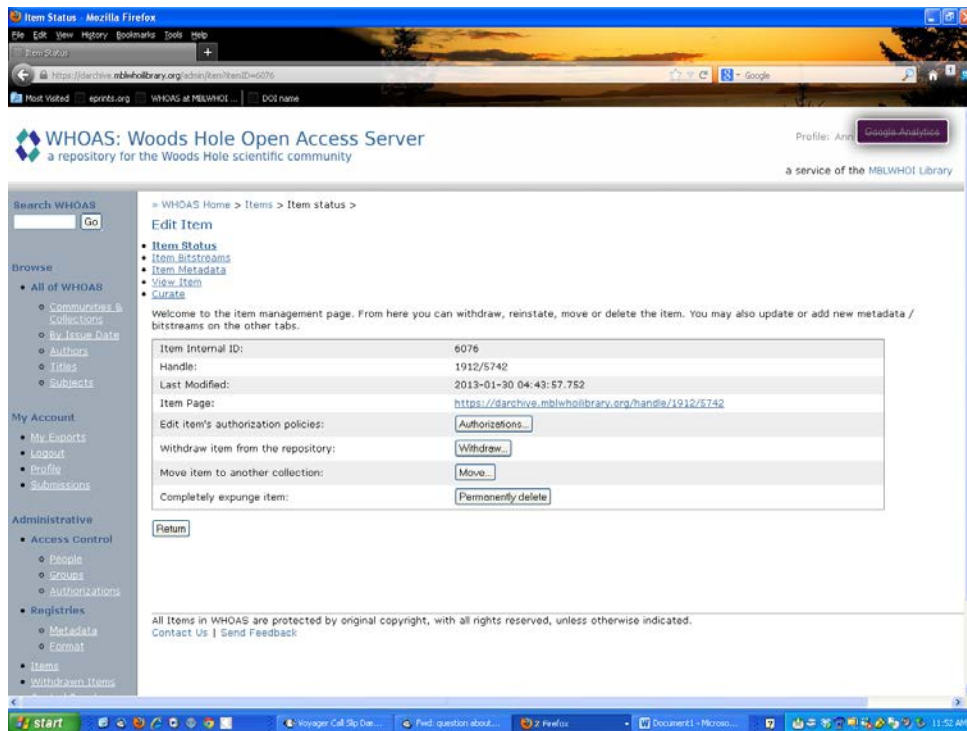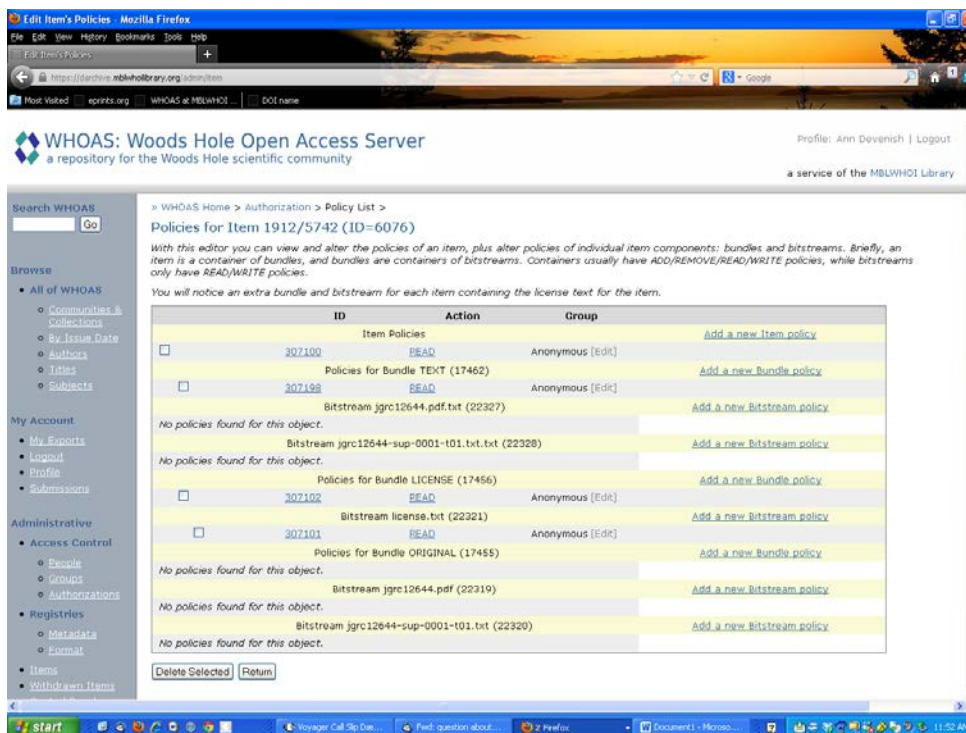
5) Save the group

Apply READ permissions to embargoed file(s)/ bitstream(s).

1) Retrieve the record with embargoed file(s)/ bitstream(s).
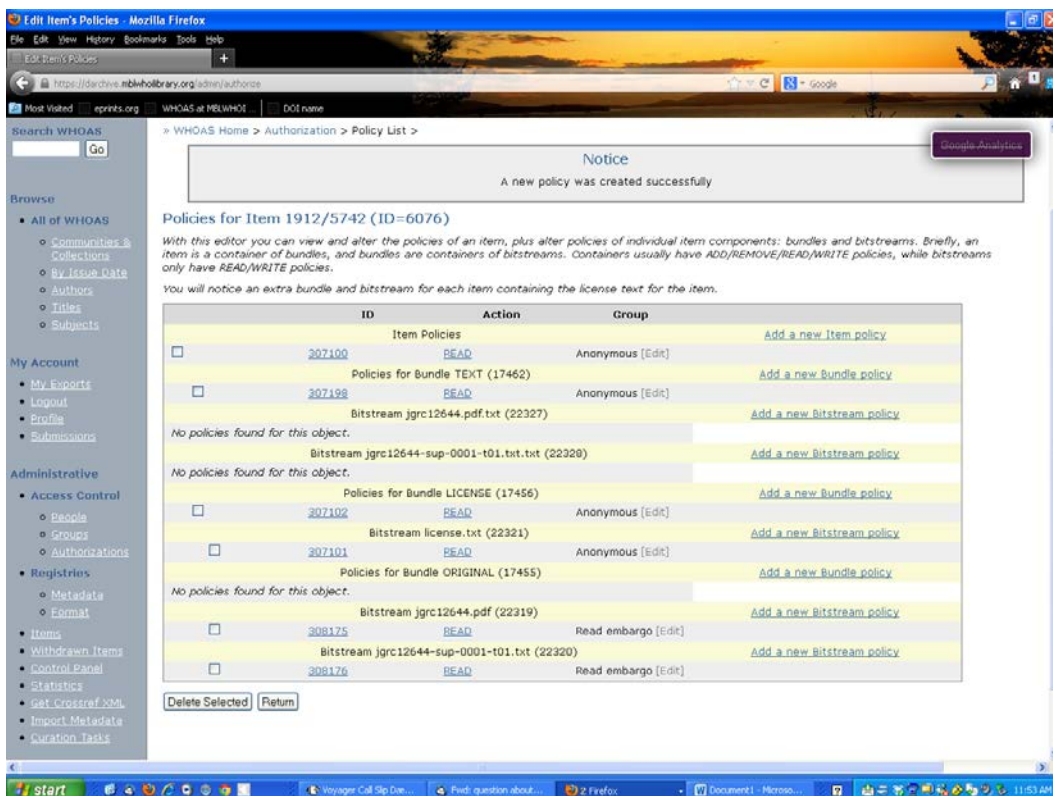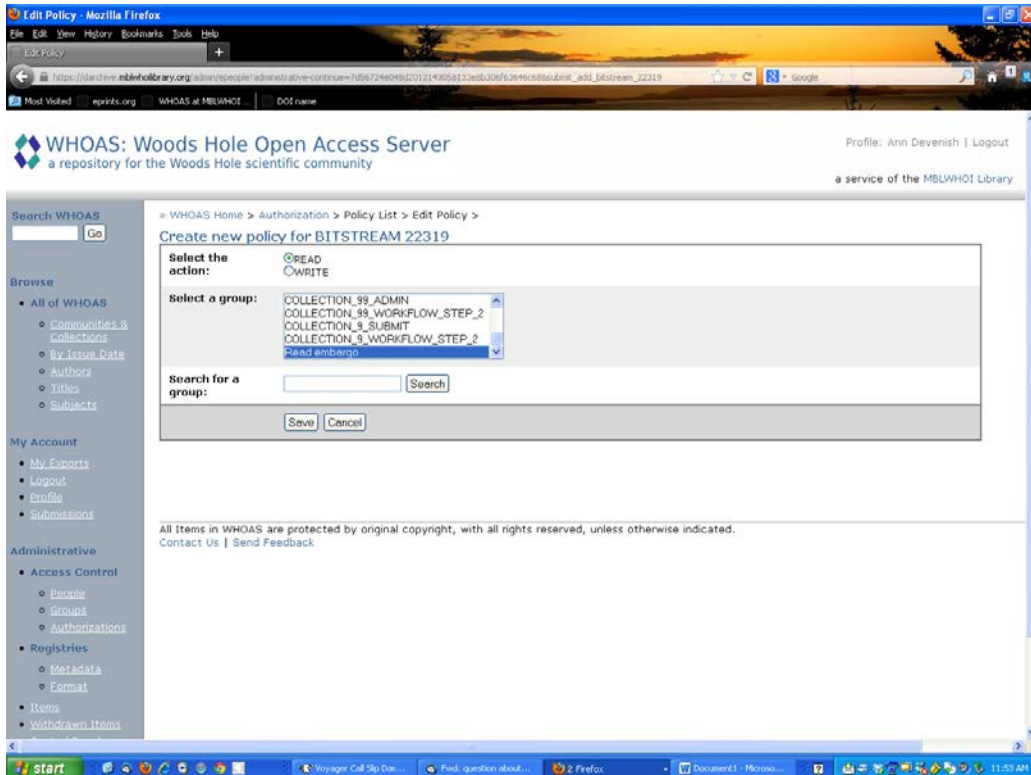
2) Select "Edit this item"

3) Select "Authorizations"

4) For the embargoed bitstreams, select "Add a new Bitstream policy"
5) Select the action READ
6) Select the previously created/edited group
7) Save
8) Repeat for each Bitstream

2.

BODC Published Data Library (including example DataCite XML record)

When a dataset author requests a DOI through BODC, they are supplied with a spreadsheet in which they fill in the required metadata fields. This is shown below, with two completed examples for published datasets:

Dataset title	Nutrient concentration profiles from long term time series at Station L4 in the Western English Channel from 2000 to 2012.	Nutrient concentration profiles from long term time series at Station E1 in the Western English Channel from 2002 to 2012.
Dataset creators	E. Malcolm S. Woodward; Carolyn Harris; Lisa Al-Moosawi.	E. Malcolm S. Woodward; Carolyn Harris; Lisa Al-Moosawi.
Dataset subject	Environment	Environment
Dataset abstract	Results are reported for the analysis of water samples taken from the L4 long-term sampling station (50°15.0'N; 4°13.0'W) situated of the south-west coast of England, United Kingdom. The analyses are reported for Nitrate+Nitrite, Nitrite, Silicate and Phosphate concentrations. The data set begins in 2000 and since 2007 has been on a more regular almost weekly sampling protocol. A classical picture is seen of generally summer nutrient depletion and with higher winter of the Nitrate, Silicate and Phosphate, however particularly in wet summers like 2012 there is a big influence of the River Tamar and spikes of high nutrient loads are seen being introduced into this near-shore sampling site.	Results are reported for the analysis of water samples taken from the E1 long-term sampling station (50°02.6'N; 4°22.5'W) situated of the south-west coast of England, United Kingdom. The analyses are reported for Nitrate+Nitrite, Nitrite, Silicate and Phosphate concentrations. The data set begins in 2002, and since 2007 has been on a more regular almost fortnightly sampling protocol, however due to the more 'offshore ' nature of the site there are many sampling weeks lost due to bad weather. A classical picture is seen of generally summer nutrient depletion and with higher winter of the Nitrate, Silicate and Phosphate.
Dataset period	January 2000 - December 2012	June 2002 - December 2012
Dataset spatial coverage	English Channel	English Channel
Dataset file format	Comma Separated Value (CSV) files.	Comma Separated Value (CSV) files.
Dataset language	English-United Kingdom	English-United Kingdom
Dataset discovery metadata record	https://www.bodc.ac.uk/data/information_and_inventories/edmed/report/732/	https://www.bodc.ac.uk/data/information_and_inventories/edmed/report/732/
Dataset publisher	British Oceanographic Data Centre	British Oceanographic Data Centre
Dataset publication date		
Dataset digital object identifer (DOI)		
Dataset citation text		
Access the dataset		

Following completion of the metadata to an acceptable standard, a DOI is minted at the DataCite metadata store. The DOI suffix is represented by a GUID, generated within a Python programming language interactive environment, as shown below:

Once the DOI is known, this DataCite XML record can be created, an example of which is below:

```
<resource          xmlns="http://datacite.org/schema/kernel-2.2"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://datacite.org/schema/kernel-
    2.2          http://schema.datacite.org/meta/kernel-
    2.2/metadata.xsd">

    <identifier identifierType= "DOI">
        10.5285/a931a96d-f08d-4e7d-af30-866f5e3e8fd8
    </identifier>

    <creators>
        <creator>
            <creatorName>
                Hennige, S.
```

```
                    </creatorName>
                </creator>
            <creator>
                    <creatorName>
                            Wicks, L.
                    </creatorName>
                </creator>
        </creators>

        <titles>
            <title>
                    Short-Term Responses of the Cold Water Coral
                    Lophelia Pertusa to Ocean Acidification
            </title>
        </titles>

        <publisher>
            British     Oceanographic    Data     Cente,     Natural
            Environment Research Council
        </publisher>
        <publicationYear>
                2012
        </publicationYear>

    </resource>
```

This record is uploaded to the DataCite Metadata Store using the sequence described in the main body of the document. This process completes the minting of the DOI. The metadata are then entered into an HTML landing page which corresponds to the URL associated with the DOI on minting. The URL of the landing page takes the form http://www.bodc.ac.uk/data/published_data_library/catalogue/{GUID}/ where {GUID} is the suffix used in the DOI-minting stage. The HTML landing page also contains links to the datafiles themselves, stored on web-accessible storage on BODC's servers.

The HTML landing page is human-readable, but the source code of the page also contains machine-readable (Resource Description Frame Work – in – attributes, or RDFa) metadata and uses the hAtom microformat to reflect changes to the page. The example lines below show an example of how this is encoded, and the following image shows the resulting machine-readable metadata.

```html
<h6>
        Published Data Library (PDL)
</h6>
<h1>
        D325 Nitrogen fixation from bioassay experiments
</h1>
<section                                    id="publicationMetadata"
                xmlns:dc="http://purl.org/dc/elements/1.1/"
                xmlns:gml="http://www.opengis.net/gml"
                xmlns:v="http://rdf.data-vocabulary.org/#"
                about="http://www.bodc.ac.uk/data/published_data_
                library/catalogue/landing_bk.html#mandfmb.html">
        <table id="Table_6">
        <!-- Required opening tags for the hAtom feed -->
                <div class="hfeed">
                        <div class="hentry">
                                <tr>
                                        <th>
                                                Dataset title
                                        </th>
                                        <td>
        <!-- Dataset title goes here -->
                                                <span    class="entry-title"
                                                        property="dc:tit
                                                        le">
                                                Nitrogen        fixation
                                                        from     bioassay
                                                        experiments    by
                                                        stable-isotope
                                                        mass
                                                        spectrometry    on
                                                        UKSOLAS    cruise
                                                        D325
                                                </span>
                                        </td>
                                </tr>
                                <tr>
                                        <th>
                                                Dataset creators
                                        </th>
                                        <td>
        <!-- Dataset creator goes here -->
                                                <span
property="dc:creator">
                                                        Dr. A. P. Rees
                                                </span>
        <!-- Other contributors go here -->
                                                ;<span
property="dc:contributor">
                                                        L Al-Moosawi.
                                                </span>
                                        </td>
                                </tr>
        <!-- Publication date and update date -->
                                <div class="published">
                                        <span        class="value-title"
                                                title="2012-01-17">
                                        </span>
                                </div>
                                <div class="updated">
                                        <span        class="value-title"
                                                title="2012-01-17">
                                        </span>
                                </div>
        <!-- Close the hAtom tags -->
                        </div>
                </div>
        </table>
</section>
```

SPARQLer Query Results

b	c
\<http://purl.org/dc/elements/1.1/title\>	" Nitrogen fixation from bioassay experiments by stable-isotope mass spectrometry on UKSOLAS cru D325 " @en
\<http://purl.org/dc/elements/1.1/description\>	" A series of bioassay type experiments were determined to investigate the impact of iron, zinc, cobal copper and nitrogen on rates of Nitrogen fixation. Experiments were performed from near surface wat using \<sup xmlns="http://www.w3.org/1999/xhtml" xml:lang="en"\>15\</sup\>N Methodology at 6 different oceanographic stations in the vicinity of the Cape Verde Islands (north-east Atlantic) betwee 13 November and 18 December 2007 onboard RRS Discovery. Stations were selected from remotely sensed information on surface chlorophyll concentrations to give a gradient of conditions between 0.0 and 0.38 µg L\<sup xmlns="http://www.w3.org/1999/xhtml" xml:lang="en"\>-1\</sup\>. The experimen was largely inconclusive, though at stations A there was evidence of limitation by Zn, Co and Cu. Interestingly Cu seemed to play a limiting role at other stations whilst nitrogen additions largely decreased rates below natural levels. " ^^\<http://www.w3.org/1999/02/22-rdf-syntax-ns#XMLLitera
\<http://purl.org/dc/elements/1.1/publisher\>	\<http://www.bodc.ac.uk/\>
\<http://purl.org/dc/elements/1.1/creator\>	" Dr. A. P. Rees " @en
\<http://purl.org/dc/elements/1.1/identifier\>	\<http://dx.doi.org/10.5285/b4d08055-8c3f-45b2-a7c6-15f8382261b3\>
\<http://purl.org/dc/elements/1.1/contributor\>	" L Al-Moosawi. " @en
\<http://purl.org/dc/elements/1.1/subject\>	\<http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_TopicCategoryCode_bic
\<http://www.opengis.net/gmlid\>	\<http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_TopicCategoryCode_bic
\<http://purl.org/dc/elements/1.1/Period\>	"start=2007-11-17; end=2007-12-13; scheme=W3C-DTF" @en
\<http://purl.org/dc/elements/1.1/coverage\>	\<http://vocab.nerc.ac.uk/collection/C19/current/SVX00015/\>
\<http://purl.org/dc/elements/1.1/format\>	"text/plain" @en
\<http://purl.org/dc/elements/1.1/language\>	\<urn:ietf:params:language:en-GB\>
\<http://purl.org/dc/elements/1.1/relation\>	\<https://www.bodc.ac.uk/data/information_and_inventories/edmed/report/4269/\>
\<http://purl.org/dc/elements/1.1/bibliographicCitation\>	" Rees, A.; Al-Moosawi, L. (2011): UKSOLAS programme cruise D325 N2 fixation bioassay experiments. British Oceanographic Data Centre - Natural Environment Research Council UK. " @en

[end]