To: SCICOM
From: DIG, PUBCOM

Introduction to Digital Citation – Towards an ICES policy on Data Citation

Introduction

Citation of journal publications is a well-accepted practice to give due credit to scientific work done by scientists, and also to signpost where others can find this information. In a similar way, citation of data can give proper credit to data providers who have made data available to the scientific community, and also provides a mechanism for tracing back scientific knowledge to the data that underpins it. However, where citation of journal publications is well established, citation of data is still in its infancy, due to important differences in the character of data as compared to publications. Despite this, major advances in the field of data citation have been made in recent years, which are highly relevant to ICES since ICES depends on data submissions by member countries to achieve its mission.

This document is a combined PUBCOM/DIG action to inform and provide advice to ICES on digital citation. As the topic evolves rapidly, updates containing information about the new developments in data citation are to be expected in the coming year(s). As a result, this document should be treated as the first of a series or as a dynamic and living document.

Digital citation, text and data

If one wishes to refer to a journal publication and cite this, one can do this in a well-established way referring to title, author name, journal name, publication date, etc. This reference will not change over time; the citation will always be correct and refer to that particular journal publication.

The question now arises how to cite a dataset. Even though the processes of publishing a journal publication and publishing an online dataset are very similar, journal publications and datasets do differ in some crucial ways (and making the data available on the web –serving the data- is different than merely publishing it).

Data citation begins to address these issues by using persistent identifiers, which can deal with many of these issues, but not with all.

Digital data citation in a nutshell

One can refer to a dataset by stating the URL (universal resource locator) or Internet address at which it can be found. However, as soon as the URL changes, the citation no longer refers to a valid address and the dataset cannot be found anymore. To remedy this, one would have to correct the citation in all instances where it has appeared or has been published. Of course, this is not feasible. The

solution is to use persistent identifiers in a citation. They come in several flavours, but all have in common several main characteristics. A persistent identifier (also called a uniform resource identifier, or URI) looks like a web address, and can take several forms. It contains the name of the computer site, such as http://dx.doi.org/ followed by a unique number or set of numbers that uniquely identifies the data. A DOI (Digital Object Identifier) is an example of a URI. Together, this address will take you to a 'landing page', which is usually a web application that resolves the URI to a specific resource in a catalogue. The landing page contains additional metadata describing the dataset and a direct link to the data, to allow the user to access and download the data. The data are referenced and stored in a way that reflects the content of that instance of the dataset at the time of the 'publication'.

The elegance of this concept is that if the location of the data changes, all one has to do is change the link only in the landing page, the resolving directory is then updated. The citation, containing a persistent identifier, still refers to the correct location where the data can be found. If the URI has been used consistently when citing the dataset, then the dataset originator should need to do nothing more. The organization issuing the persistent identifier is responsible for maintaining the link between the persistent identifier and the actual data in the repository.

Reasons to digitally cite data

By having persistent identifiers included in a citation, data citation will be used to give due credit to data providers. In addition, it is increasingly important to have a direct link between the results of a scientific study and the underlying data, which can be made explicitly with a persistent identifier. Furthermore, the landing page may also contain some text on how the data provider wishes the data to be cited or referenced. A system of citation metrics and associated impact factors, similar to the one in use for the more traditional journal publications, can and should be developed by a joint effort of the scientific community, funding agencies, data centres and journal publishers.

Digital data: citing static and non-static data

Digital citation as understood here covers data that support a published paper's tables and figures. Such data is given a DOI and the data is frozen and static, in other words this is **static data**. From a data management perspective handling such information is relatively straightforward; however a few points need to be observed:

(i) ensuring that the DOI link address is always available, which can be delegated to a service provider.
(ii) changes in the data will not be reflected in the cited data as the cited data is now disconnected from its original source. There are mechanisms that allow new versions of DOI to be cited but this is something that the author of the paper must manage.

The vast majority of data are subject to change: new data get added, existing data are modified or deleted. In addition the model that the data represents is subject to change over time. These features give rise to a number of crucial differences when managing **non-static data** for DOI purposes:

(i) the physical location of the data can change so the citation will reference a location which does not exist at that address (or link) anymore;
(ii) over time the structure, layout or organisation of the data can change. This can affect some DOI schemes that use hierarchical pathways to

describe their datasets in the same way as a scientist might reorganise the folder structure of their hard-drive or email. For example, the columns in the data file are all right-shifted by one position because a new column has been introduced in the left-most position for some, typically technical, reason;

(iii) the semantics, meaning, or intention around a piece of data may change, which may also be reflected in the address or link to the digital data. For example, the business meaning of "fishing effort" may evolve over time to encompass new concepts or deprecate others;

(iv) the actual value of a piece of data may change. For example, a typographic error needs to be corrected or a new version of an existing value needs to be adapted.

Challenges to be addressed

Two issues concerning the use of persistent identifiers need to be resolved:

(a) the role and funding of the persistent identifiers' issuing authorities. Each persistent identifier costs a small amount of money, but with the amount of datasets available, this may easily grow to a huge total sum for each data contributor. For explanation - The data originator/holder pays a small one-time fee for the assigned «data handle», or URI (Uniform Resource Identifier). The fees for data handles vary depending on the volume of handles requested, the country in which the assigning agency is located, and how the certified/authorized repositories are funded.

(b) When it's out there, someone has to manage it. **Temporality**, or how data changes over time, in terms of data existence at a representational and at a physical level can play havoc with data sets. Changing structure and changing semantics imply a changing schema or a "schema evolution". In terms of exposing such data through a digitally cited interface, such a solution is at the edge or possibly beyond the state of the art at the moment.

Digital data citation systems

There are a number of initiatives minting and managing persistent identifiers. Two of the most commonly referred to are DataCite (http://www.datacite.org/) and CrossRef (http://www.crossref.org/). Both of these use Digital Object Identifiers (DOIs) as their persistent identification. The DOI system has become an ISO standard (ISO 26324:2012) which sets it apart from other persistent identifiers. ICES already successfully uses DOIs for ICES Journal of Marine Science articles (via CrossRef). DataCite is aimed at the data networks, while CrossRef is aimed at publishing networks. More information on CrossRef can be found in the annex to this document, and the ICES publications department would recommend joining CrossRef so that other ICES publications can gain similar exposure as the ICES Journal of Marine Science.

For discussion within ICES

As long as the originators of the data get the authorship credit for their efforts, because they use persistent identifiers for their data, the system will work. However, if ICES is going to add persistent identifiers to data at the regional dataset level, and on the national level (where the data originated) no persistent identifiers are going to be used, then ICES will get the acknowledgement for digital citation and it is more difficult to give the correct digital

acknowledgement to the contributors on the national level. One option is for ICES to assign persistent identifiers and also list the author or data contributor. Also, ICES can be listed as another site having the data. As an example, see how BCO-DMO[1] is mentioned at this DOI: http://dx.doi.org/10.1575/1912/5075.

More information

The Amsterdam Manifesto on Data Citation Principles:
http://www.force11.org/AmsterdamManifesto

Ocean Data Publication Cookbook:
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10574

http://ands.org.au/cite-data/index.html

http://wokinfo.com/products_tools/multidisciplinary/dci/

Ball, A. & Duke, M. (2012). 'How to Cite Datasets and Link to Publications'. DCC How-to Guides.

Edinburgh: Digital Curation Centre. Available online:
http://www.dcc.ac.uk/resources/how-guides

Ball, A., Duke, M. (2012). 'Data Citation and Linking'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online:
http://www.dcc.ac.uk/resources/briefing-papers/

Davidson, J. (2006). "Persistent Identifiers". DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre. Handle: 1842/3368. Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

Lowry, R., E. Urban, and P. Pissierssens (2009), A New Approach to Data Publication in Ocean Sciences, Eos Trans. AGU, 90(50), 484–484, doi:10.1029/2009EO500004.

SCOR/IODE Workshop on Data Publishing, Oostende, Belgium, 17-19 June 2008. Paris, UNESCO, 23pp. 2008. (IOC Workshop Report No. 207) (English) [http://www.scor-int.org/Publications/wr207.pdf]

CODATA-ICSTI Task Group (editor Yvonne M. Socha) "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data"
https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article

1 The Biological and Chemical Oceanography Data Management Office (Woods Hole, US)

Annex 1: Further information on a citation system for publications: CrossRef

CrossRef

Association of scholarly publishers representing 4000 publishers

DOI registration agency – one most suited to general publications in the sciences

March, 2013: 90m clicks on CrossRef DOI links

Membership fee

Based on annual publishing revenue e.g. basic fee = $275 p.a.

Fee for depositing new DOI = $1 (for content 2011-2013)

Fee for depositing new DOI for archive content (pre-2011) = $0.15

E.g. 2012: (excluding EG reports, advice sheets – no ISBN) we published 20 documents = $295

Condition of membership

18 months after joining, must link all references used in your material that you have submitted (only if they have an existing DOI).

Minimum metadata need by CrossRef to assign a DOI to a journal article/publication:

- Journal title;
- ISSN; (either print or online or you can provide both)
- first author;
- year;
- volume and issue page numbers;
- the URL where the content is located.

Note: CrossRef does not require members to assign DOIs to their historical archive if they do not wish to; we can start from the next publication and work forward. Therefore, no significant employee/financial resources needed as was the case with MSS and CM documents projects.

Annex 2: Further information on a citation system for publications; DataCite

DataCite

Global non-profit organisation representing information managers (Libraries and Data Centres).

25+ organizations involved as national nodes for assigning DOIs.

Membership

Membership is open to all not for profit organisations who wish to allocate DOI names. A member should be actively working with data centres for the purpose of issuing DOIs. The membership fee for full members is €8,500 p.a.

Alternatively, if you are only interested in getting DOIs minted, you can work with the national node and negotiate a price with them directly.

The national allocation agency for DataCite in Denmark is DTU. An exploratory enquiry was made and ICES was offered DOI numbers for datasets.

Pricing

DTU will charge a flat-rate of €1350 Euros p.a. regardless of the number of DOIs. They can also sell them separately for €680 Euros/piece.

Naming

We will actually get our own Datacentre ID similar to '1093' in '/10.1093/'. After the last '/' we will have full control (anything that can be used in a url). They advise being careful with the use of names in the DOI, because they cannot be changed again., but it could look like ices.di[0-9]

Persistence

A DOI should link to a persistent entity and any change should in theory result in a new DOI. The ICES datasets are dynamic, but it does not make sense to constantly create new DOIs for our datasets. They suggested versioning (e.g. /1, /2, /3, ...).

DataCite Metadata

When creating a DOI it is mandatory to supply some metadata. It should be an ISO standard, but it is not clear if it is the same as ISO 19139/19115 that ICES currently uses.