Data Publication in Oceanography: OGS-NODC experience

Matteo Vinci, Alessandra Giorgetti, Alberto Brosich and Alessandro Altenburger

OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) Borgo Grotta Gigante 42/C - 34010 - Sgonico (TS) - Italy

Abstract: The amount of available oceanographic data is rapidly growing thanks to the autonomous measuring systems and to the increasing number of accessible repositories. The relationship between source data and other resources like scientific publications becomes essential. The challenge for the scientific institutions is to develop a strategy to ensure its long-term data accessibility with a clear acknowledgement of data originators.

Keywords: long term data access, acknowledgements, metadata, permanent identifiers

1. Introduction

EU is requesting to increase the open access to source data and to all publications related to them aiming to realize the as widest as possible linked and open data scenario. Without any doubt this will increase the circulation of key information in all sectors of knowledge and is a hopeful objective but there is the need to take care of different needs of the actors involved. From one side there is the interest of people who want to access to source data and related publications to take advantage of the existing high level information. From the other side there is the interest of the data originators that would like to be properly acknowledged for their efforts in order to be rewarded with founding that will allow them to continue with their activity. Between these actors the data centers have the chance to play a relevant role focusing on providing access to data, track them and clearly attribute credit to providers to encourage data dissemination.

2. Persistent identifiers and data publication

The publication of data on a website, as has been done for projects outcomes for a lot of years, is a relatively easy task and can be considered a first step to encourage data accessibility. This makes the information accessible but without any long-term commitment there are no guarantees about long term availability

Received: 24.6.2016 Accepted: 21.8.2016 ISSN 2241-1925

or that the files haven't become corrupted. 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, by ensuring that the dataset is complete, frozen and has enough supporting metadata, allows to use data in the years to come. Publishing also allows data producers to obtain academic credit for their work in creating the datasets**.** (Leadbetter, A et al: 2013)

All the previous aspects can be safeguarded by an adequate assignment and management of persistent identifiers. There is a quite wide panorama of permanent identifiers that can be used to tag different objects like electronic documents but also scientists.

An identifier system can be considered in different ways by different communities that have specific needs. For example for some communities the provision of a unique name to an object will be considered a key point, for some others the focus could be more shifted in the set of services that allow the linkage between the unique name to the resource or to the set of metadata provided. Specific requirements will differ, but it is vital that institutions seeking to assign permanent identifiers to datasets recognize 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. (Leadbetter, A et al: 2013)

Between the available permanent identifiers there are the Digital Object Identifiers -DOIs-. They are represented by a character strings used to uniquely identify an object that is stored with a set of relevant metadata stored as XML - eXtensible Markup Language - document following the schema provided by the issuing authority. All the efforts related to the DOIs minting and management are focused on the long term re-usability of the information. For example the set of metadata associated to the object provides users with all the needed information to understand which data they are handling also long time after data creation. The information provided is related to the creator, the title, the publisher, the period, the description of the document etc. This supply the as clearer as possible picture of the resource depending on the set of information provided by the data originator.

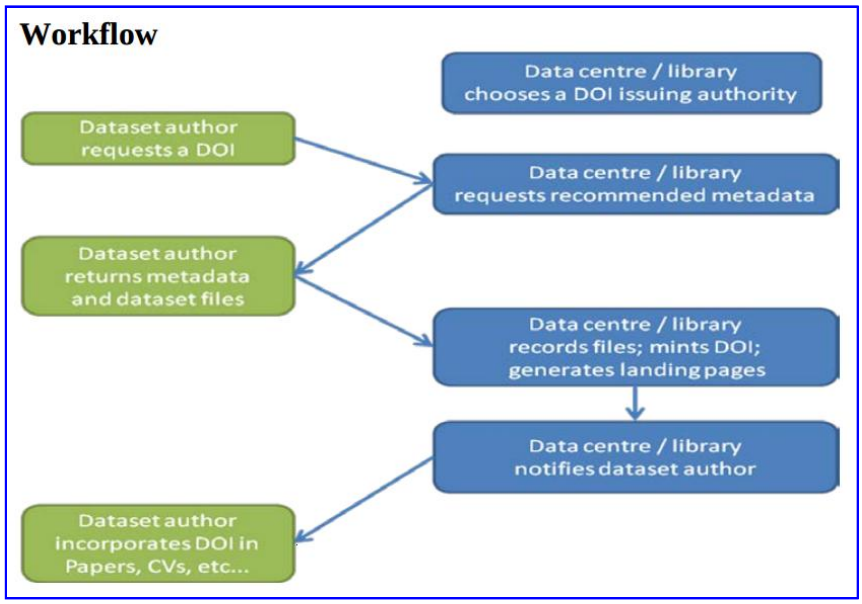
Furthermore there is a commitment between the institute who is minting DOIs and the DOI provider to maintain active and updated a "landing page" where all the DOIs should be kept available.

A data centre that want to able to manage persistent identifiers like DOIs need first to choose an issuing authority (DataCite in OGS NODC case) and then need to provide the landing page where all the digital documents can be retrieved.

When a dataset authors make the request for a new DOI the data centre ask first for the document and the relevant metadata. When all the needed information is delivered the data centre mints the DOI and make it available through the

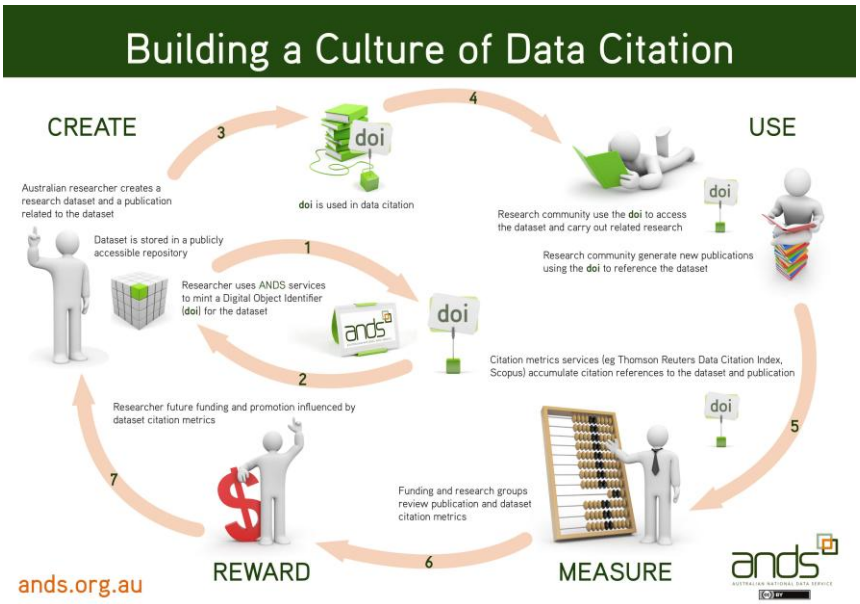
landing page. Now the document is ready to be included in CVs or publications by the author.

Fig. 1- DOI minting workflow description from "Ocean Data Publication Cookbook" (IOC Manuals and Guides 64, Version 1, March 2013)



To mint a DOI OGS-NODC follow the "Ocean Data Publication Cookbook" guidelines (IOC Manuals and Guides 64, Version 1, March 2013).
DOIs have to follow the same format: a prefix (assigned through DataCite) 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 (a 32-character hexadecimal string).
The assignment of a DOI enables accurate data citation because each DOI name permanently and unambiguously identifies the object to which it is associated.
The dataset authors receive full citation credit for their data collection efforts because are clearly listed in the DOI metadata, furthermore if the DOI is linked to the permanent identifier for scientists the relevant metrics can be directly updated. Data publication enabling an adequate data citation can certainly be an incentive to make data accessible because the citation of high quality datasets provides the evidence of the quality of work done by the data originator.

Fig. 2- benefits of the data citation from "Building a culture of Data Citation" poster of Australian National Data Service (ANDS)



3. OGS –NODC and the oceanographic data management experience

OGS-NODC is a National Oceanographic Data Centre in the network of institutions of International Oceanographic Data Exchange –IODE- of IOC UNESCO. Its role is the active collection, quality check, standardization and dissemination in a long term perspective. Furthermore there is a focus to act in support to national data originators, promoting data sharing following EU policies and standards and provide data visualization and products useful for EU policies (e.g. MSFD, MSP).

Between these activities, data publication should be considered the hat of a work done following the principle of "Capture once and use many times".

Two main Oceanographic data management communities provide driving principles for data publication activities: IODE and the Research Data Alliance.

The "cookbook for data publication" provided by the IODE supply an overview of the best practices available in the community that can be used as starting points from the data centers willing to manage permanent identifiers. The set of recommendations for the data citation of evolving data provided by RDA face the use of a static identifier to tag a set of evolving information, a more complex but common situation.

4. DOI use cases

The minting and citation of a permanent identifier could be linked to two main use cases: static datasets, evolving datasets.

In the first case the data centre is dealing with a static set of information that can be a report or a set of historical measurements. In this situation the management is quite simple. There is the need to collect all the needed information to mint the DOI and load it in the landing page.

In the case of an evolving dataset we have to use a static identifier to tag an evolving set of information. There are not specific rules about how to deal with this case but a set of best practices and recommendations.

Two concrete examples about this case are the Argo floats network and the Seismic networks.

Argo floats are a network of floating autonomous buoy that collects high-quality temperature and salinity profiles from the upper layers to2000m of the ice-free global ocean and currents from intermediate depths. The data come from battery-powered autonomous floats that spend most of their life drifting at depth where they are stabilised by being neutrally buoyant at the "parking depth" pressure by having a density equal to the ambient pressure and a compressibility that is less than that of sea water. When reach the surface the instrument deliver the information to a data collecting centre by satellite communication. For this reason the collected dataset is continuously growing.

The suggested behavior to manage this situation foresees to mint a DOI with a general description of the dataset. Then will be minted other DOIs for the snapshots of the datasets in continuous evolution with a granularity agreed with the originator. For the Argo the case the agreed granularity is of one month.

For the seismic networks the situation is similar. Usually there is available a network of seismographs in specific area for a specific period. A general DOI will be minted for the description of the monitoring network and then related DOIs for snapshots with a given granularity will be associated to the first one.

For both of the described cases the XML tag used for the linkage of DOIs is <relatedIdentifiers>.

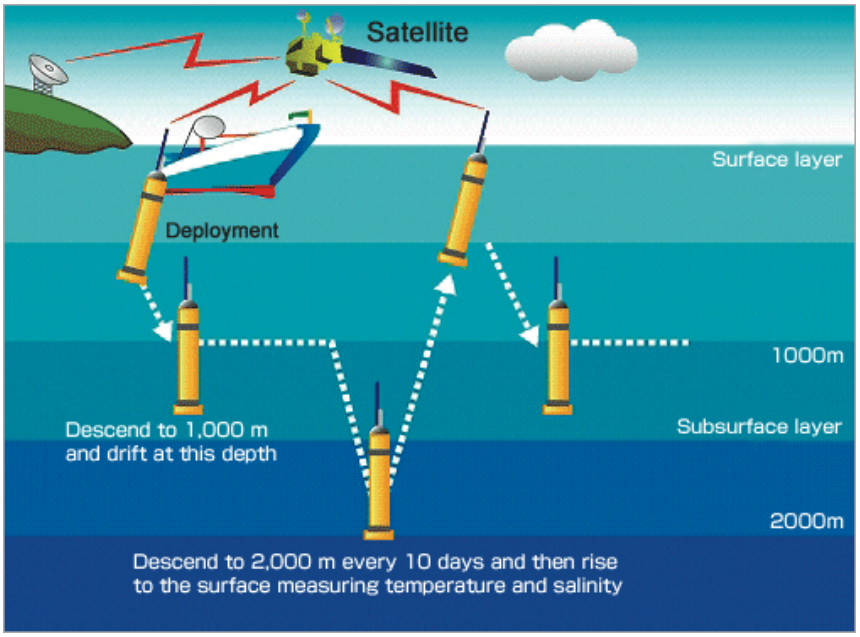
The last recommendations from RDA are suggesting a more “query centric” point of view for the management of DOIs foreseeing the great effort on the management of evolving datasets.

The prerequisites for this approach are a database with the timestamped versioning of data and a query store facility where archive the queries done to the database and the metadata describing what the queries are requesting.

The principle is to assign a DOI for each query to the database to be able to retrieve with the versioning information the needed version of available data.

The interest of the user could be focused on retrieving the version of the queried dataset used in a specific moment or to obtain the last version of it to see if calibrations or corrections caused significant changes to the data.

Fig 3. Argo floats working cycle



5. Conclusions

The growth and circulation of available oceanographic information is increasing bringing up the need of the following points: long-term accessibility, usage track and clear credit attribution to providers. A proper use of permanent identifiers could safeguard the previous aspects providing the chances to: publish data giving them visibility, acknowledge providers and save the necessary set of metadata able to identity source data even long time after their measurement.

References

Leadbetter, A., Raymond, L., Chandler, C., Pikula, L., Pissierssens, P., Urban, E. (2013) Ocean Data Publication Cookbook. Paris: UNESCO, 41 pp. & annexes. (Manuals and Guides. Intergovernmental Oceanographic Commission, 64), (IOC/MG/64)

Rauber, A., Asmi, A., van Uytvanck, D. and Pröll S. Data Citation of Evolving Data Recommendations of the Working Group on Data Citation (WGDC)

IODE - The programme "International Oceanographic Data and Information Exchange" (IODE) of the "Intergovernmental Oceanographic Commission" (IOC) of UNESCO http://www.iode.org/

DataCite https://www.datacite.org/

ANDS - Australian National Data Service (ANDS) National Collaborative Research Infrastructure Strategy (NCRIS). http://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation