

# A single DOI for Argo; a generic approach to making datasets that grow and evolve with time citable on legacy infrastructure

**Justin Buck**, National Oceanography Centre (United Kingdom), [juck@bodc.ac.uk](mailto:juck@bodc.ac.uk)

**Thierry Carval**, IFREMER (France), [thierry.carval@ifremer.fr](mailto:thierry.carval@ifremer.fr)

**Thomas Loubrieu**, IFREMER (France), [thomas.loubrieu@ifremer.fr](mailto:thomas.loubrieu@ifremer.fr)

**Frederic Merceur**, IFREMER (France), [frederic.merceur@ifremer.fr](mailto:frederic.merceur@ifremer.fr)

## Abstract

The Argo dataset grows and evolves with time and changes in the expectations on the citation of Argo data and traceability of data citations has driven a 5 year effort to make Argo data citable via a single DOI. This has now been implemented by Ifremer on the Argo dataset for the first time using an approach that enables citation for the Argo data at monthly snapshots without requirement for significant enhancement to the Argo data infrastructure. The approach presented is readily applicable to other data infrastructures and enables Argo to partly meet the recommendations of the Research Data Alliance Dynamic Data Citation working group.

## Paper

Unambiguous citation of data used in academic publications is crucial for the transparency and reproducibility of science especially when results are used as evidence to underpin national and international policy. Data citation of static datasets is well established and documented. However, the citation of dynamic data is an area of active development. When data are cited it should also be possible to unambiguously resolve the data. A key milestone in this development was when DataCite defined three methods for citing dynamic data; cite times slices, snapshots, or the dataset with an access date. This thinking was progressed by the RDA data citation working group in 2015 with the publication of an ideal method for citing dynamic data. These recommendations have not been implemented on real data yet and full implementation is not possible for legacy data systems where there are insufficient resources to upgrade software and infrastructure. In addition to this the syntax for including dynamic data information has not been agreed or ratified with DataCite and CrossRef. The IPCC 5<sup>th</sup> assessment used evidence from data collected from observing programmes, high impact reports such as this need the data that underpin contributing papers to be unambiguously citable.

To address outstanding dynamic data citation issue a single DOI has been implemented on the Argo dataset that enables long term resolution of previous versions of the Argo dataset with a monthly resolution. This has enabled the citation of Argo data through a single DOI (example: <http://dx.doi.org/10.17882/42182>). This DOI is associated to a default Landing Page. By default, this landing page would presents the last monthly snapshot with links for displaying all the previous monthly snapshots. Each of these snapshots will be associated to a unique key.

In this default landing page, it is possible to find:

- the general metadata
- a “How to cite” section
- a “Related publications” section automatically filled with Megan Scanderbeg bibliographic survey

some links to export the metadata in 3 different formats including the RIS format to include the citation in software like Endnote For dynamic data citation especially, a keystore is managed by the publisher

infrastructure (Ifremer/Seanoë) so to handle snapshot citation for this same enriched DOI (example: <http://dx.doi.org/10.17882/42182#41950>). Then the landing page will be automatically adapted to offer metadata and data to the snapshot requested:

- The title, the date and the citation is adapted.
- The data section presents the snapshot as default but the link will still be available to list the others snapshots.

This solution is ready applied to other legacy data systems where the amount of available storage does not constrain the storage of monthly snapshots. A significant advantage of this solution is that is an economic way to implement dynamic data citation on legacy datasets where resource is insufficient for the full implementation of the RDA dynamic data recommendations.