RCN Open Data Working Group
Data Publication/Data Citation Task Team

San Francisco, California
December 2, 2012
Data Publication/Citation

The Task Team considered models for “data publication/data citation” as a way to create incentives for researchers to share their data.
Incentive

• Data Publication that enables accurate data citation can indeed be an incentive for researchers to make data available.

• Data citation metrics are starting to be collected. Will they be considered in performance assessments?

• Funding agency and publisher mandates.
Use Cases

1. Data related to traditional journal articles are assigned persistent identifiers referred to in the articles and stored in institutional repositories.

2. Data held by data centers are packaged and served in formats that can be cited.
Data Citation

The goal of the use cases has been to identify best practices for *tracking data provenance and clearly attributing credit to data creators/providers so that researchers will make their data accessible*. The assignment of persistent identifiers, specifically Digital Object Identifiers (DOIs), enables accurate data citation.
Digital Object Identifier (DOI)

- A **digital object identifier (DOI)** is a character string used to uniquely identify an object. Metadata about the object is stored in association with the DOI name.
- Libraries have been using for years, now de facto standard for data.

http://dx.doi.org/10.1575/1912/5105
10 - DOI registry
1575 - DOI registry agent – CrossRef
1912 - publisher – MBLWHOI
5105 – “item” number
Use Case One

- MBLWHOI Library has worked directly with authors as well as with a data management office to deposit data and metadata.
- DOIs have successfully been assigned to data associated with articles before and after publication of the article.
Science Direct Links to Associated Data in the Woods Hole Open Access Server (WHOAS)
Collaboration with NSF funded Biological and Chemical Oceanography Data Management Office (BCO-DMO)

- Automated the ingestion of metadata from BCO-DMO for deposit, with a copy of each dataset into the Institutional Repository WHOAS.
- Incorporates functionality for BCO-DMO to request a DOI from the Library.
- Partnership allows the Library to work with a trusted data repository to ensure high quality data while BCO-DMO utilizes the library services and is assured a permanent copy of the data is associated with the DOI.
Use Case Two

- **The Published Data Library (PDL)** is implemented by the British Oceanographic Data Centre. It provides snapshots of specially chosen datasets that are archived using rigorous version management. The publication process exposes a fixed copy of an object and then manages that copy in such a way that it may be located and referred to over an indefinite period of time. Using metadata standards adopted across NERC's Environmental Data Centres, the repository assigns DOIs obtained from the British Library/DataCite to appropriate datasets.
# Sample Record from PDL

## AMT14 Rates of nitrogen fixation in surface waters

**Dataset title**: Rates of nitrogen fixation in surface waters by stable-isotope mass spectrometry during the AMT programme cruise AMT14

**Dataset creators**: Dr. A. P. Rees; N. Milward.

**Dataset subject**: Biota

**Dataset abstract**: The RRS James Clark Ross sailed from Port Stanley in the Falkland Islands on 29 April and arrived in the UK on 1 June 2004. The main objective of the cruise was to determine the trophic state of the planktonic ecosystem with respect to large-scale patterns in the ambient concentrations and supply to the surface waters of inorganic nutrients and of dissolved organic matter. Nitrogen fixation using 15N methodology was determined at all stations sampled with rates ranging between 0.4 and 8.8 mmol L\(^{-1}\) d\(^{-1}\) in the South Atlantic Gyre, between 0.2 and 2.2 mmol L\(^{-1}\) d\(^{-1}\) in the North Atlantic Gyre and reached maximum rates of 17.0 mmol L\(^{-1}\) d\(^{-1}\) within the area of equatorial upwelling.

**Dataset period**: 3rd May 2004 - 22nd May 2004

**Dataset publisher**: British Oceanographic Data Centre

**Dataset publication date**: 17th January 2012

**Dataset digital object identifier (DOI)**: doi:10.5285/41479c42-4d1b-4da9-be97-4e532ce19922

**Access the dataset**: Dataset — [amt14_nfw_dataset](#)

**Metadata**: [amt14_nfw_description.html](#)
Other Similar Models

- Dryad
- Pangea
- Dspace@MIT

Other major projects also investigating linked data/enhanced publication include DRIVER and OPenAIREplus.

This type of repository is meant to function in addition to national and subject data centers, not as a replacement.
Peer Review

• Community acceptance of an agreed peer review procedure for data publication is due.

• Data publication in data repositories ensure provenance, attribution, permanence and quality metadata. Does not guarantee the scientific quality of the data.
Conclusion

The Task Team concludes that Data Publication that enables Data Citation can certainly be an incentive to make data accessible. The associated functionality to deposit data safely and securely should be attractive to the researcher and of course the additional citation of the data associated with a research paper will add value to these data as an essential component of research output.