2010 International Digital Curation Conference, Chicago, IL   Poster #165

Enhancing The Recognition, Reusability, And Transparency Of Scientific Data Using Digital Object Identifiers

Environmental
Data Science & Systems

Bruce Wilson (wilsonbe@ornl.gov), Tammy Beaty, Robert B Cook, Christopher Lenhardt, Jon Grubb,
Les Hook & Carol Sanderson

ORNL DAAC started using DOIs in 2007

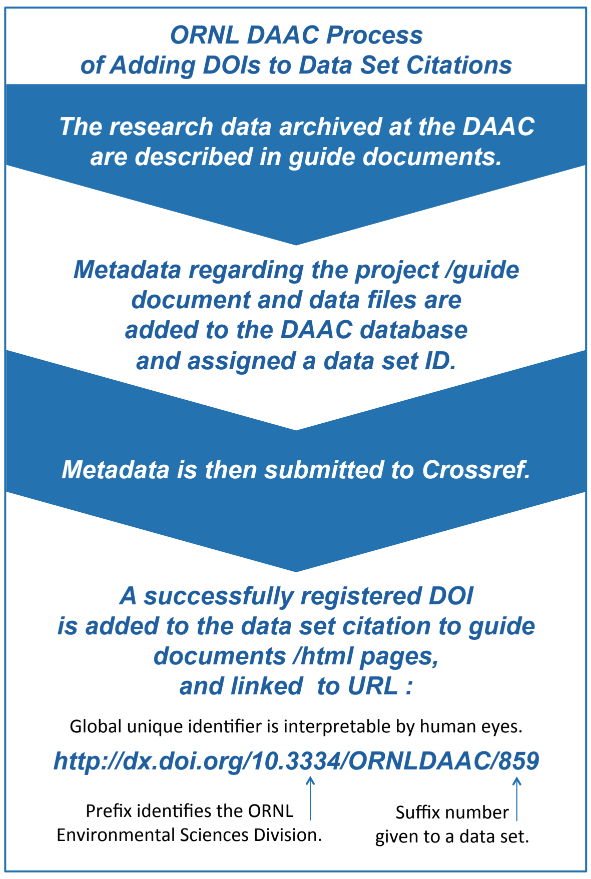
The Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC), part of the NASA Earth Science Data and Information System (ESDIS) project, is responsible for archiving and distributing a wide range of terrestrial ecology data sets. Partly to enhance the recognition for scientists sharing their data, the ORNL DAAC has had a data citation policy for many years, with the citation in the name of the scientists who collected and providing an Internet URL pointing to the data set. Some journal editors, however, objected to a URL in a scientific citation, arguing that URLs are transient and problematic for the anticipated lifetime of a scientific journal article. In response to this concern, the ORNL DAAC started assigning Digital Object Identifiers (DOIs) to published data sets in 2007 and incorporating the DOI in the requested citation for each data set. DOIs have now been assigned to all ORNL DAAC published data sets.

Data Citation (credit and attribution)
vs.
Data Identification (reproduce and reuse)

These are two separate problems, and we are dealing primarily with the data citation need.

Example Data Set Citation:

Wofsy, S. C., S. R. Saleska, E. H. Pyleand, L. R. Hutyra. 2008. LBA-ECO CD-10 Tree DBH Measurements at the km 67 Tower Site, Tapajos National Forest. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. *doi:10.3334/ORNLDAAC/859*



ORNL DAAC Process of Adding DOIs to Data Set Citations

The research data archived at the DAAC are described in guide documents.

Metadata regarding the project /guide document and data files are added to the DAAC database and assigned a data set ID.

Metadata is then submitted to Crossref.

A successfully registered DOI is added to the data set citation to guide documents /html pages, and linked to URL :

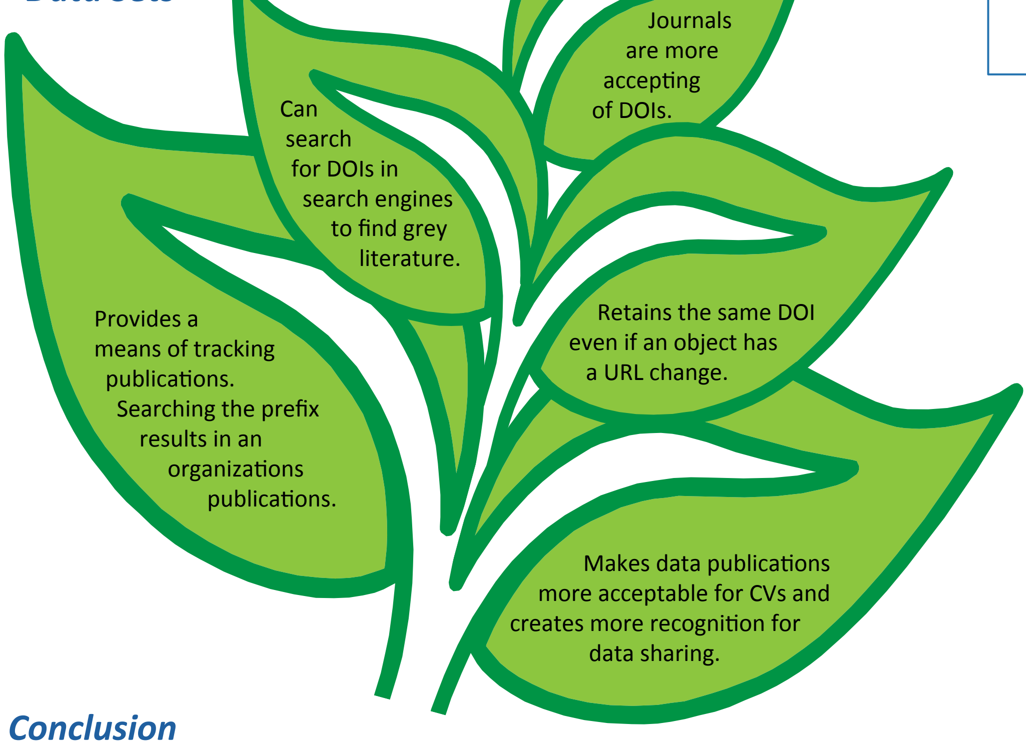
Global unique identifier is interpretable by human eyes.

http://dx.doi.org/10.3334/ORNLDAAC/859

Prefix identifies the ORNL Environmental Sciences Division.

Suffix number given to a data set.

Benefits of Adding DOIs to Finalized Data Sets



Journals are more accepting of DOIs.

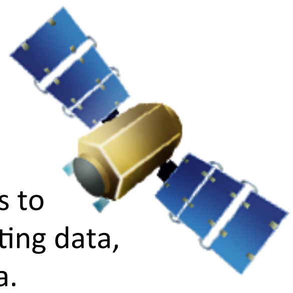
Can search for DOIs in search engines to find grey literature.

Provides a means of tracking publications. Searching the prefix results in an organizations publications.

Retains the same DOI even if an object has a URL change.

Makes data publications more acceptable for CVs and creates more recognition for data sharing.

Future Work



- Assign identifiers to constantly updating data, like satellite data.

- Assign identifiers to individual data files

Conclusion

DOIs are very useful for finalized data sets, as well as data sets that are updated infrequently,. DOIs also work well for for managing data set citations. We have not assigned DOIs to dynamically generated data sets, such as those generated by our data subsetting tools (such as the MODIS subsetting tool and the dynamic subsets generated by OGC web services). Dynamic data sets may be a case where separating data set identification (for scientific reproducibility) from data set citation (for attribution and impact analysis) may be appropriate. DOIs have also improved our ability to track citations of data sets, both in the formal scientific literature and in documents published to the general Web. We are now seeing examples where researchers are listing published data sets on their CV, as one indication of improved recognition of the value for sharing and archiving data sets. DOIs are not yet useful for tracking and assessing impact in other applications, such as use in the educational setting or in decision support applications. For those types of applications, the resulting document(s) are often not as easy to locate and identify as the peer-reviewed scientific literature, making the impact assessment more difficult. Currently, the DOI resolves to a document that describes the data set and provides a link to the data set itself. We are evaluating approaches to enable more direct machine access to the underlying data, in support of tools such as scientific workflow engines.

Oak Ridge National Laboratory is operated for the Department of Energy by UT-Battelle under contract number DE-AC05-00OR22725
The ORNL DAAC is funded by the National Aeronautics and Space Administration under interagency agreement NNG09HP12I

OAK RIDGE
National Laboratory

NASA