Geophysical Research Abstracts
Vol. 18, EGU2016-11639-1, 2016
EGU General Assembly 2016

How do you assign persistent identifiers to extracts from large, complex, dynamic data sets that underpin scholarly publications?

Lesley Wyborn (1), Nicholas Car (2), Benjamin Evans (3), and Jens Klump (4)

(1) National Compuational Infrastructure, Australian National University, ACT, Australia (lesley.wyborn@anu.edu.au), (2) Geoscience Australia, Canberra, ACT, Australia (nicholas.car@ga.gov.au), (3) National Compuational Infrastructure, Australian National University, ACT, Australia (ben.evans@anu.edu.au), (4) Australian Resources Research Centre, CSIRO, Perth, WA, Australia (jens.klump@csiro.au)

Persistent identifiers in the form of a Digital Object Identifier (DOI) are becoming more mainstream, assigned at both the collection and dataset level. For static datasets, this is a relatively straight-forward matter. However, many new data collections are dynamic, with new data being appended, models and derivative products being revised with new data, or the data itself revised as processing methods are improved. Further, because data collections are becoming accessible as services, researchers can log in and dynamically create user-defined subsets for specific research projects: they also can easily mix and match data from multiple collections, each of which can have a complex history. Inevitably extracts from such dynamic data sets underpin scholarly publications, and this presents new challenges.

The National Computational Infrastructure (NCI) has been experiencing and making progress towards addressing these issues. The NCI is large node of the Research Data Services initiative (RDS) of the Australian Government's research infrastructure, which currently makes available over 10 PBytes of priority research collections, ranging from geosciences, geophysics, environment, and climate, through to astronomy, bioinformatics, and social sciences. Data are replicated to, or are produced at, NCI and then processed there to higher-level data products or directly analysed. Individual datasets range from multi-petabyte computational models and large volume raster arrays, down to gigabyte size, ultra-high resolution datasets. To facilitate access, maximise reuse and enable integration across the disciplines, datasets have been organized on a platform called the National Environmental Research Data Interoperability Platform (NERDIP). Combined, the NERDIP data collections form a rich and diverse asset for researchers: their co-location and standardization optimises the value of existing data, and forms a new resource to underpin data-intensive Science.

New publication procedures require that a persistent identifier (DOI) be provided for the dataset that underpins the publication. Being able to produce these for data extracts from the NCI data node using only DOIs is proving difficult: preserving a copy of each data extract is not possible due to data scale. A proposal is for researchers to use workflows that capture the provenance of each data extraction, including metadata (e.g., version of the dataset used, the query and time of extraction). In parallel, NCI is now working with the NERDIP dataset providers to ensure that the provenance of data publication is also captured in provenance systems including references to previous versions and a history of data appended or modified.

This proposed solution would require an enhancement to new scholarly publication procedures whereby the reference to underlying dataset to a scholarly publication would be the persistent identifier of the provenance workflow that created the data extract. In turn, the provenance workflow would itself link to a series of persistent identifiers that, at a minimum, provide complete dataset production transparency and, if required, would facilitate reconstruction of the dataset. Such a solution will require strict adherence to design patterns for provenance representation to ensure that the provenance representation of the workflow does indeed contain information required to deliver dataset generation transparency and a pathway to reconstruction.