Geophysical Research Abstracts
Vol. 21, EGU2019-6377, 2019
EGU General Assembly 2019

Linking Derivatives from Big Earth and Environmental Reference Datasets to Publications is Not Easy: Perspectives from the NCI Petascale Data Repository

Lesley Wyborn and Ben Evans
Australian National University, National Computational Infrastructure, Acton, Australia (lesley.wyborn@anu.edu.au)

The recently revised Commitment Statement by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) requires research outputs related to publications to be FAIR (Findable, Accessible, Interoperable, and Reusable) and have unique persistent identifiers assigned to each generated/produced dataset. This is relatively straightforward for small/isolated datasets developed by a particular research project over a defined period: outputs are then made available for download as files that relate to a specific publication. Any repository distributing such datasets acts more as a traditional library: no changes or enhancements are made to the dataset.

More recently, some large data repositories (or data storage facilities) are now managing major reference collections (e.g., satellite Earth observation, geophysics) in response to a growing demand for research at larger geographical scales and/or over longer time periods. Data from multiple surveys and/or progressive time series acquisitions are aggregated into co-located reference collections: some are further standardised and structured into High Performance Datasets (HPD) for use in Data-Intensive Science.

To further enhance these reference datasets, advanced multi-faceted repositories are making innovative data services and tightly-managed compute environments available so that researchers can dynamically access, analyse and transform the data on-the-fly, extract subsets and then perhaps process the data streams in their own computing environment (clouds, supercomputers, portal services, or desktops) using a mixture of their own code, third party software, services and libraries. Some derived datasets can be voluminous and complex. Given the dynamic nature of this processing, it is not cost effective, or at times even possible, for each derivative dataset to be saved. So how is this type of data referenced? Connections between data as published by the repository and the derivative datasets as used in the publication are hard to make FAIR-compliant. In particular, what should the persistent identifier reference to enable effective citation and reproducibility?

One approach is to autogenerate and publish the references to all dependent datasets and the ‘recipe’ on how the derivative dataset is created. But many users do not appreciate all the reference datasets used and are not readily able to publish third party workflows themselves. In addition, data storage/repository facilities are constantly being upgraded and optimised, particularly when new or replacement hardware and software infrastructure (e.g., component operating systems, libraries, software and services). Thus, to help support persistent citation, repositories now need to record and make publicly accessible (preferably in standards compliant, machine query-able ways) any changes to any dataset, as well as ensuring that changes to the data services software that provide data access are also recorded.

Supporting consistent approaches to citation for derivatives from large reference datasets requires community agreement and consideration of costs and technical solutions. Developing such publication standards will not be easy: they have to be consistent across multiple Earth and environmental centres internationally and include not just research repositories but also government repositories. Increasing dependency by researchers on commercial cloud data storage services such as Google, Amazon, etc., means they also need to be engaged in this process.