Geophysical Research Abstracts
Vol. 19, EGU2017-12087, 2017
EGU General Assembly 2017

Improving global data infrastructures for more effective and scalable analysis of Earth and environmental data: the Australian NCI NERDIP Approach

Ben Evans, Lesley Wyborn, Kelsey Druken, Clare Richards, Claire Trenham, Jingbo Wang, Pablo Rozas Larraondo, Adam Steer, and Jon Smillie

Australian National University, National Computational Infrastructure, ACTON, Australia (ben.evans@anu.edu.au)

The National Computational Infrastructure (NCI) facility hosts one of Australia's largest repositories (10+ PBytes) of research data collections spanning datasets from climate, coasts, oceans, and geophysics through to astronomy, bioinformatics, and the social sciences domains. The data are obtained from national and international sources, spanning a wide range of gridded and ungridded (i.e. line surveys, point clouds) data, and raster imagery, as well as diverse coordinate reference projections and resolutions.

Rather than managing these data assets as a digital library, whereby users can discover and download files to personal servers (similar to borrowing 'books' from a 'library'), NCI has built an extensive and well-integrated research data platform, the National Environmental Research Data Interoperability Platform (NERDIP, http://nci.org.au/data-collections/nerdip/). The NERDIP architecture enables programmatic access to data via standards-compliant services for high performance data analysis, and provides a flexible cloud-based environment to facilitate the next generation of transdisciplinary scientific research across all data domains.

To improve use of modern scalable data infrastructures that are focused on efficient data analysis, the data organisation needs to be carefully managed including performance evaluations of projections and coordinate systems, data encoding standards and formats. A complication is that we have often found multiple domain vocabularies and ontologies are associated with equivalent datasets. It is not practical for individual dataset managers to determine which standards are best to apply to their dataset as this could impact accessibility and interoperability. Instead, they need to work with data custodians across interrelated communities and, in partnership with the data repository, the international scientific community to determine the most useful approach. For the data repository, this approach is essential to enable different disciplines and research communities to invoke new forms of analysis and discovery in an increasingly complex data-rich environment.

Driven by the heterogeneity of Earth and environmental datasets, NCI developed a Data Quality/Data Assurance Strategy to ensure consistency is maintained within and across all datasets, as well as functionality testing to ensure smooth interoperability between products, tools, and services. This is particularly so for collections that contain data generated from multiple data acquisition campaigns, often using instruments and models that have evolved over time. By implementing the NCI Data Quality Strategy we have seen progressive improvement in the integration and quality of the datasets across the different subject domains, and through this, the ease by which the users can access data from this major data infrastructure. By both adhering to international standards and also contributing to extensions of these standards, data from the NCI NERDIP platform can be federated with data from other globally distributed data repositories and infrastructures.

The NCI approach builds on our experience working with the astronomy and climate science communities, which have been internationally coordinating such interoperability standards within their disciplines for some years. The results of our work so far demonstrate more could be done in the Earth science, solid earth and environmental communities, particularly through establishing better linkages between international/national community efforts such as EPOS, ENVRIplus, EarthCube, AuScope and the Research Data Alliance.