Geophysical Research Abstracts
Vol. 18, EGU2016-5362, 2016
EGU General Assembly 2016

Unleashing Geophysics Data with Modern Formats and Services

Alex Ip (1), Ross C. Brodie (1), Kelsey Druken (2), Irina Bastrakova (1), Ben Evans (2), Carina Kemp (1), Murray Richardson (1), Claire Trenham (2), Jingbo Wang (2), and Lesley Wyborn (2)

(1) Geoscience Australia, Canberra, ACT, Australia, (2) National Computational Infrastructure, The Australian National University, ACT, Australia

Geoscience Australia (GA) is the national steward of large volumes of geophysical data extending over the entire Australasian region and spanning many decades. The volume and variety of data which must be managed, coupled with the increasing need to support machine-to-machine data access, mean that the old "click-and-ship" model delivering data as downloadable files for local analysis is rapidly becoming unviable – a "big data" problem not unique to geophysics. The Australian Government, through the Research Data Services (RDS) Project, recently funded the Australian National Computational Infrastructure (NCI) to organize a wide range of Earth Systems data from diverse collections including geoscience, geophysics, environment, climate, weather, and water resources onto a single High Performance Data (HPD) Node. This platform, which now contains over 10 petabytes of data, is called the National Environmental Research Data Interoperability Platform (NERDIP), and is designed to facilitate broad user access, maximise reuse, and enable integration. GA has contributed several hundred terabytes of geophysical data to the NERDIP.

Historically, geophysical datasets have been stored in a range of formats, with metadata of varying quality and accessibility, and without standardised vocabularies. This has made it extremely difficult to aggregate original data from multiple surveys (particularly un-gridded geophysics point/line data) into standard formats suited to High Performance Computing (HPC) environments. To address this, it was decided to use the NERDIP-preferred Hierarchical Data Format (HDF) 5, which is a proven, standard, open, self-describing and high-performance format supported by extensive software tools, libraries and data services. The Network Common Data Form (NetCDF) 4 API facilitates the use of data in HDF5, whilst the NetCDF Climate & Forecasting conventions (NetCDF-CF) further constrain NetCDF4/HDF5 data so as to provide greater inherent interoperability.

The first geophysical data collection selected for transformation by GA was Airborne ElectroMagnetics (AEM) data which was held in proprietary-format files, with associated ISO 19115 metadata held in a separate relational database. Existing NetCDF-CF metadata profiles were enhanced to cover AEM and other geophysical data types, and work is underway to formalise the new geophysics vocabulary as a proposed extension to the Climate & Forecasting conventions. The richness and flexibility of HDF5's internal indexing mechanisms has allowed lossless restructuring of the AEM data for efficient storage, subsetting and access via either the NetCDF4/HDF5 APIs or Open-source Project for a Network Data Access Protocol (OPeNDAP) data services. This approach not only supports large-scale HPC processing, but also interactive access to a wide range of geophysical data in user-friendly environments such as iPython notebooks and more sophisticated cloud-enabled portals such as the Virtual Geophysics Laboratory (VGL).

As multidimensional AEM datasets are relatively complex compared to other geophysical data types, the general approach employed in this project for modernizing AEM data is likely to be applicable to other geophysics data types. When combined with the use of standards-based data services and APIs, a coordinated, systematic modernisation will result in vastly improved accessibility to, and usability of, geophysical data in a wide range of computational environments both within and beyond the geophysics community.