The Australian Geoscience Data Cube – Progress and Future Directions

Alex Ip, Dr Robert Woodcock

Geoscience Australia, Canberra, Australia, alex.ip@ga.gov.au
CSIRO, Canberra, Australia, robert.woodcock@csiro.au

ABSTRACT

The Australian Geoscience Data Cube (AG-DC) is a common analytical framework for large volumes of regular gridded geoscientific data initially developed by Geoscience Australia and now being developed in collaboration with CSIRO and the National Computational Infrastructure (NCI) Facility at the Australian National University (ANU).

Since its development in early 2013, the AG-DC has already enabled high-resolution (25m) continental-scale, multi-decadal analysis of the entire available Landsat archive (over 10^{13} data points) via a prototype API. Users from Geoscience Australia (GA) and CSIRO and collaborators are now utilising the massive co-located HPC, cloud and storage resources of the NCI to prepare data collections and conduct new data-intensive science on them.

In this BoF session we will be describing both existing analysis environments and proposed access mechanisms to the AG-DC. In particular, work is continuing to provide access to seamless spatio-temporal subsets of the data via web services and ultimately to provide standard interfaces for in-situ Petabyte-scale processing at the NCI. Work is also continuing to both simplify and extend the interface through which users at the NCI currently gain access to entire data collections for in-situ, whole-of-collection processing.

We will share some of the use cases currently exploiting the AG-DC and seek to communicate with others interested in performing large-scale, cross-domain analyses of regular gridded geoscientific data, whether that data already exists in the AG-DC or not. We are particularly interested to discuss how we might address any novel use cases and also to determine which new geoscientific datasets might be prioritised for inclusion in the common analytical framework. Additional discussion around geospatial considerations and other potential interoperability issues such as cross-calibration will also be encouraged.

EXTENDED ABSTRACT

Birds of a Feather Presentation – 20 minutes

Brief history of the AG-DC

The Unlocking the Landsat Archive (ULA) project completed at the end of June 2013 was a $3.5M initiative funded through the Australian Space Research Program. The fundamental aim of the project was to improve access to Australia's archive of Landsat data, and provide an analysis capability for delivery of environmental information to inform and support government policy. The project generated around 0.4PB of Landsat 5 & 7 data from some 240,000 individual scenes captured over a fifteen year period (1998 through 2012). The data collection is now part of NCI's National Environmental Research Data Collections, with storage partly funded by the Research Data Storage Initiative (RDSI).

In early 2013, it became clear to the National Earth Observation group (NEO) at GA that there were no third-party tools available or even nearing completion anywhere in the world which would allow GA to meet operational commitments including the Murray-Darling Basin (MDB) vegetation analysis and the National Flood Risk Information Portal (NFRIP).

The difficulties arose largely because higher-resolution remote sensing data, such as that produced by Landsat, is often spatially and temporally sparse and irregular (unlike model output). Consequently, it is not well suited to manipulation using the monolithic multi-dimensional array constructs typically employed to manage large-scale gridded data such as climate model output.

NEO developed a working prototype of the raster data management system now at the core of the Australian Geoscience Data Cube (AG-DC). The design met the scientific and operational requirements and it has since been successfully employed to produce several high-resolution, multi-decadal, continental-scale analyses.

The collaboration between GA, CSIRO and the NCI extended the AG-DC as a common analytical framework for a wide range of data collections. As part of this collaboration, two hands-on training workshops were conducted in early 2014 for scientists from various stakeholder agencies and institutions. Many of these researchers are now already developing other analyses to run against the AG-DC at NCI using the existing Application Programming Interface (API).

Some AG-DC use cases

- Land-cover analysis and monitoring (e.g. forestry, cropping & pasture growth monitoring)
- Water detection and temporal analysis of inundation
- Tidal-zone bathymetry and monitoring the health of remote water bodies
- Burn Scar and other disaster recovery analysis
- MODIS-Landsat data fusion
- Urban growth monitoring
- Identification of calibration/validation sites for Earth Observation data
- MODIS-Landsat blending (which has applications in all of the above)

Potential issues in data interoperability

Geometry – ortho-rectification and standardised Discrete Global Grid Systems must be employed to ensure pixel-on-pixel conformance between individual datasets.
Geodesy – The coordinate reference system used must satisfy as many of the Goodchild-Kimerling criteria [1] as possible, and consideration should also be given to the accommodation of dynamic datums.
Cross-calibration and adjustment – Remote sensing data should be corrected to a standard surface-reflectance to permit interoperability between disparate data collections. Similarly, elevation and bathymetry data must be levelled and adjusted to a consistent reference frame.
Collection maintenance and versioning – It is not practical to duplicate extremely large data collections in their entirety, so some mechanism must be provided for managing changes and identifying both dataset and collection level versioning.

Current and proposed access mechanisms

Access to the data is currently provided via a Python API which queries the metadata database and presents the temporally-arranged paths to the underlying data files which fall in the spatio-temporal range requested. Work is underway to both simplify and extend the interface through which users at the NCI currently gain access to entire data

collections for in-situ, whole-of-collection processing. This work will create opportunities for new access modes and easier support for languages other than Python.

Work is also continuing to provide access to seamless spatio-temporal subsets of the data via OpenDAP and OGC web services (e.g. WCS) which provides a first stage of API service access to the data for users outside the NCI. This mode of access will support the use of subsets of the data collection in Virtual Laboratory environments, but it may not be practical for very large subsets or entire data collections.

Ultimately, the goal is to extend the protocols ro use web processing services that provide standard interfaces for in-situ Petabyte-scale processing at the NCI (e.g. WPS, WCPS).. There are open issues associated with the delivery of these services, including identity management and provenance capture.

Future directions for the AG-DC

The current version of the AG-DC uses data subdivided into spatially-regular, time-stamped 2D tile files managed by a relational database. This model provides the flexibility required for highly dynamic collections but is inherently I/O intensive. Proof-of-concept work has already been undertaken for the reorganization of the data into multidimensional data structures which aggregate the observations into spatial partitions, thereby extending the concept of data partitioning from the spatial domain to the temporal. The parameterized nature of the AG-DC will permit the testing of a wide range of parameter sets (e.g. spatial and temporal partition sizes, coordinate reference systems, etc) against important use cases in specific environments in order to provide good performance across a range of common use cases with a single data store.

Work is continuing to develop a framework which will abstract away the management of parallel processing for various workflows and access modes. It is highly desirable that such a management framework will be applicable not just to the NCI, but also to cloud environments including public infrastructure.

Birds of a Feather Discussion – 40 minutes

Potential use cases for the AG-DC

Session participants will be invited to propose potential new use cases for the data. We will discuss the avenues currently available to those wishing to implement and apply new analysis algorithms against the AG-DC holdings, and also discuss access methods including those actively under development and those planned for the future.

Suggestions for sustainable community contributions of formal datasets

The strategic goal for the AG-DC is to provide a common analytical framework for large-scale data collections which will enable cross-domain data-intensive science. Several new data collections are already being processed for inclusion in the AG-DC, but the AG-DC collaborators are also seeking to hear from the custodians of other regular gridded datasets which might be of broad interest to the research community. These data custodians will find out how they can be assisted to add their collections to the AG-DC and join the expanding scientific community sharing data in this framework.

Potential issues in data interoperability

Session participants will be invited to discuss the interoperability issues raised in the presentation with a view to identifying potential solutions. These issues include coordinate reference and gridding system selection, radiometric and geometric correction as well as quantifying and representing uncertainty in position, time and values.

Any questions posed by session participants

Session participants will be invited to pose general questions to the AG-DC collaborators.

REFERENCES

1. Michael F. Goodchild and A. Jon Kimerling (editors) – Discrete Global Grids: A Web Book (http://ncgia.ucsb.edu/globalgrids-book/) - Ch. 8 (Keith C. Clarke) Criteria and Measures for the Comparison of Global Geocoding Systems.

ABOUT THE SPEAKERS

Alex Ip is Senior Data Analyst in High Performance Data at Geoscience Australia. Alex is a computer systems engineer with over twenty-eight years of industry experience in software development and data management. Alex has developed distributed job processing systems for genetic analysis, extensive databases for commercial and industrial applications, real-time motion control systems for industrial robotics and, most recently, petabyte-scale raster data management systems.

Dr Robert Woodcock is Stream Leader – National Geoscience Information Infrastructure and Integration, for the CSIRO Minerals Down Under Flagship and Deputy Director Informatics for the CSIRO Earth Observation Informatics Future Science Platform. He is responsible for a portfolio of activities involving national spatial data infrastructure and Earth observation data analysis and integration platforms. His portfolio has seen application in Industry and Government including the AuScope Australian geoscience information infrastructure and national and international Earth science information exchange and environmental information platforms.