

# **Bringing Australian Geophysical Data onto a High Performance Data Node at the National Computational Infrastructure\***

**Jingbo Wang<sup>1</sup>, Irina Bastrakova<sup>2</sup>, Ben Evans<sup>1</sup>, Carina Kemp<sup>2</sup>, Ryan Fraser<sup>3</sup>, and Lesley Wyborn<sup>1</sup>**

Search and Discovery Article #70210 (2016)

Posted February 1, 2016

\*Adapted from extended abstract prepared in conjunction with poster presentation at AAPG/SEG International Conference & Exhibition, Melbourne, Australia, September 13-16, 2015, AAPG/SEG © 2016.

<sup>1</sup>National Computational Institute, Acton, ACT, Australia ([jbw\\_cam@hotmail.com](mailto:jbw_cam@hotmail.com))

<sup>2</sup>Geoscience Australia, Canberra, ACT, Australia

<sup>3</sup>CSIRO, Perth, WA, Australia

## **Abstract**

The National Computational Infrastructure (NCI) at the Australian National University (ANU) has organised a priority set of 30+ large volume national earth and environmental data assets on a High Performance Data (HPD) node within a High Performance Computing (HPC) facility, as a special node under the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS) Research Data Storage Infrastructure (RDSI) project. The Australian National Geophysical Collection was identified as a nationally significant collection and approved as one of the RDSI funded collections. It includes the most comprehensive publicly available collections of Australian airborne magnetic, gamma-ray, seismic, electromagnetic, magnetotelluric, and gravity data sets. The total size allocated for this geophysical data collection is currently 300 Terabytes. Organising this major geophysical data collection within a high performance computing environment creates a new capacity for accessing and processing data at both high resolution, and at full-continent spatial extent. Further by co-locating and harmonising the geophysical data assets with other significant national digital data collections (e.g., earth observation, geodesy, digital elevation, bathymetry) new opportunities have arisen for Data-Intensive interdisciplinary science at a scale and resolution not hitherto possible. To support this integrated HPC/HPD infrastructure our data management practices include co-development of Data Management Plans (DMP) with the data collection custodians; the development of standards compliant catalogues on data collections/data sets; and minting and maintaining persistent identifiers. The data are accessible either via direct access or via international standards compliant data services including geospatial standard (ISO 19115) catalogues, metadata harvesting protocols (OAI-PMH), and OGC protocols. A Virtual Geophysics Laboratory has also been established that links the geophysical data assets with online software and tools using cloud based scientific workflows.

## Introduction

National Computational Infrastructure (NCI) at the Australian National University (ANU) is an integrated High Performance Computing and Data (HPC-HPD) environment comprising a 1.2 PetaFlop supercomputer (Raijin), a HPC class 3000 core OpenStack cloud system (Tenjin), and several highly connected large-scale high-bandwidth Lustre filesystem. NCI operates as a formal partnership between the ANU and three of the major Australian National Scientific Agencies: the Commonwealth Scientific and Industrial Research Organisation (CSIRO), the Bureau of Meteorology (BoM) and Geoscience Australia (GA) who are also the custodians of many large volume national scientific data collections. The data from these national agencies and collaborating overseas organisations are either replicated to, or produced at, NCI: in many cases they are processed to higher-level data products. Model data from computational workflows at NCI are also captured and released as modelling products. Observational data is typically captured by instruments and then processed either at NCI or at one of the agencies, to produce data products that are then made available on the facility. The data can be accessed directly through high-performance access, and the highly skilled users generally prefer this method. Alternatively, by establishing Virtual Laboratories, which are cloud based scientific workflow portals, data, software, and computational environments can be managed through a single interface. Virtual Laboratories ultimately help not only to use the data more effectively and efficiently, they also makes online software and applications, which are already enabled to run on that data, more accessible particularly to less experienced users.

GA, in collaboration with the State and Territory Geological Surveys is the steward for data in the Australian Geophysical Data Collection, which contains data gathered since at least 1951. Major data sets include the airborne geophysics data set containing approximately 32.8 million line kilometres of data, and the gravity data set, which holds over 1.57 million reliable onshore stations from more than 1800 surveys. The collection also includes a large number of seismic surveys from onshore explosive, wide-angle reflection and refraction surveys, as well as seismic surveys across offshore basins. Data are accessible under either Creative Commons by 4.0 license or by individual state government license agreements.

However, over time, as data volumes grow exponentially, the resolution and size of the digital geophysical data sets rapidly reach the point where they exceed the capacity of the on-premise computational infrastructure of the government agencies and their industry clients to store and dynamically access them. Processing and analysing the data to its fullest resolution is also challenging due to the scale of the computation required. Many geophysical data sets can now only be analysed if they are averaged, subsampled, or split into smaller tiles. It is also impossible to deliver the large volume data sets online: most have to be copied onto hard media and then shipped by mail.

To meet these multiple challenges, in 2011 GA began a trial to mirror some geophysical data sets to the NCI to utilise their integrated HPC/HPD facilities to enable processing of far larger volumes of data at either much higher resolution and/or over larger areal extents. At NCI it was clear that the HPD/HPC technologies not only improved the quality and resolution of scientific outputs, results were produced in much faster time frames. HPC also allowed consideration of a range of likely scenarios with uncertainties to be routinely expressed and visualized: for example, it was now possible to know how many points were used in the definition of each particular surface with quantified uncertainties for each measurement.

Under the RDSI program, NCI sought to identify research data holdings of lasting value and importance and to contribute funding to their development at the NCI. This government collection of geophysical data was nominated and accepted in 2014. The Australian Geophysics Data Collection is progressively being made accessible to the community (including research, industry and government) and is currently free of charge.

The Australian Geophysical Data Collection is available to the community either through direct access or via international standards compliant data services, including ISO 19115 compliant catalogues, and OPeNDAP and OGC protocols via the NCI Data Management Portal. The data services can also be discovered and processed online through the Virtual Geophysical Laboratory using either the cloud or HPC.

### **Data Sets in the Australian Geophysical Data Collection**

The data sets in the Australian Geophysical Collection are listed in [Table 1](#) with current and projected size and format. The coverage of some of the grid and line data are illustrated in [Figure 1](#), [Figure 2](#), [Figure 3](#), and [Figure 4](#).

### **NCI's Data Management Web Portal**

NIC has created a Data Management Web Portal that has five key initiatives to both support data management practices at NCI and to also enable federated governance arrangements to be established with the custodians/owners of the source data sets, particularly for those data sets that are exact mirrors from their sites. The five initiative are: the Data Management Plan, a metadata creation tool, a catalogue cross referencing system, a digital object identifier minting service, and a THREDDS data access portal.

### **Data Management Plan (DMP)**

The Australian Geophysical Data Collection is managed by NCI with over 30+ major data collections (see NCI's website <http://nci.org.au/data-collections/data-collections/>). A key element is NCI's Data Management Plan (DMP); where all attributes in this plan are compatible with the ISO 19115 metadata standards and hence metadata can be automatically generated from the DMP. This ensures that the metadata is then easily transferable and interoperable for sharing and harvesting. The DMP is used along with metadata from the data itself, to create a hierarchy of data collection, dataset, and time series catalogues that is then exposed through NCI's GeoNetwork catalogue (<http://geonetwork.nci.org.au>). This hierarchy of records are linked using a parent-child relationship. The top-level metadata for Collection is available from <http://geonetwork.nci.org.au/geonetwork/srv/eng/metadata.show?id=14&currTab=simple> whilst descriptions of the individual surveys within the collection is on <http://geonetworkrr2.nci.org.au>. At NCI, we are currently improving the metadata interoperability in our catalogue by linking with standardized community vocabulary services, to help harmonise data from different national and international scientific communities.

## **NCI's Metadata Creation Tool**

The metadata web portal ([Figure 5](#)) has been developed with authentication management to data managers and NCI users. It minimizes the workload for managing metadata records and exposing through the NCI catalogue service.

## **NCI-GA Catalogue Cross-Referencing**

Geoscience Australia (GA) and state government are the major data source providers for the Australian Geophysical Data Collection. Therefore, the catalogue system of the individual data providers and NCI's catalogue needs to be cross referenced and synchronised with minimum manual interaction. For example, GA's master catalogue (GeoCat) hosts all their internal catalogue entries. When the data is mirrored to NCI's data repository, the GA catalogue needs to reflect that there are now two locations of this data set. If a derived product is generated and published at NCI, this new product needs to be updated in GA's catalogue. Consistently managing and updating the two catalogues systems is critical.

## **NCI's Digital Objective Identifier (DOI) Minting Service**

Data citation is another important aspect of the NCI data infrastructure, which allows acknowledge and credit of the data producer/contributor/publisher. NCI is capable of providing Digital Object Identifiers (DOIs) minting services with support from the DataCite partnership agreement of the Australian National Data Service (ANDS). Through DOI's we can track data usage and encourage data sharing.

## **NCI's THREDDS Data Access Portal**

The published data can be accessed through NCI's THREDDS Data Access Portal. [Figure 6](#) is a screen shot for various ways of accessing a particular gravity data set.

## **The Virtual Geophysical Laboratory**

Data in the Australian Geophysical Data Collection can also be discovered and processed through the Virtual Geophysical Laboratory (VGL, [Figure 7](#)), which is a cloud-based scientific workflow portal that provides geophysicists with access to an integrated environment that connects the data, open source software and cloud/HPC computing technology. The VGL was developed as a collaborative project between CSIRO, GA, and NCI, and was funded by the Federal Government's SuperScience program through the National eResearch Collaboration Tools and Resources (NeCTAR) Project.

VGL is a paradigm change: enabling a transition from the traditional delivery of geophysical data via file download for users, to now providing users with online access to data via web-services with connections to geophysical processing and inversion tools. The VGL provides a mapping interface ([Figure 7](#)) to use spatial and other attributes to discover and filter the geophysical data resources. VGL is more an infrastructure than an application: it orchestrates the linking of data to a variety of software resources that are also available as online services. Throughout the

workflow, provenance information is collected and stored as a provenance record, thus enhancing transparency of any result and enabling spatial discovery of who ran what and when. The key benefit of this provenance workflow approach to geophysical processing is that all procedures used are accessible, verifiable and can be used by others to independently validate the results.

## **Future Work**

Our future work will be focused on 3 key areas: increasing exposure and knowledge of the existence of the Australian Geophysical Data Collection; enhancing programmatic access to the data (including authenticated access), and researching new formats to improve usage of geophysical data in massively parallel environments (Cloud, HPC).

By taking a standards based approach and using cross walks to several catalogue standards such as ISO 19115, OAI-PMH, DCAT, the current NCI geophysical catalogue entries are being automatically harvested by other catalogues, such as ANDS Research Data Australia, FIND, and the University of California at San Diego Supercomputer Centre. This is helping to increase the exposure and knowledge of the Australian Geophysical Data Collection both nationally and internationally.

However, one of the challenges to increase direct programmatic access to the data itself is the lack of an internationally agreed vocabulary and related ontologies for geophysical data that are specifically designed for machine to machine interaction. Such vocabularies exist in the geosciences (e.g. the International Union of Geological Sciences endorsed vocabularies that support GeoSciML, EarthResourceML, and Geological Time Definitions (<http://resource.geosciml.org/>)). There are some local efforts to develop equivalent geophysical data vocabularies and ontologies, but these initiatives need to be coordinated at an international level.

Modern HPC and cloud hardware trends are towards increased multicore parallelism and larger memory. The higher capacity computational infrastructures can also process much larger data volumes meaning that data from multiple geophysical surveys can be aggregated and processed over larger spatial areas. Many existing geophysical data standards, particularly those that have been around for over 20 years, are now being seriously challenged. For example, the traditional seismic data standard is SEG Y, was first developed in 1975, with a revised version published in 2002. When aggregating multiple surveys, a series of problems with the SEG Y format become apparent, such as non-standard header information (the textual header can be in ASCII or EBCDIC); performance issues with saving data; and binary data is stored in IEEE or IBM floating point. These are barriers to processing seismic data efficiently using HPC. Promising preliminary results with HDF5 format, suggest it might be reasonable to develop a new standard for the seismology community, which could replace the aging SEG Y standard. We propose to trial some data conversion from SEG Y to HDF5 and benchmark the performance by running standard migration algorithm using the widely tested case for comparison.

As these developments are taking place, VGL will be further expanded to include access to more online geophysical processing tools and algorithms. New data sets will be added as they become available, but it is clear that migration of older data sets into more modern formats that better support parallel processing is inevitable. VGL will continue to utilise more powerful compute facilities and storage volumes to vastly improve the quality, resolution and timeliness of geophysical processing. Users of VGL will access their data from NCI, do their required

processing online in situ and then simply download the results. The days of locating and downloading individual files, locally subsetting them and then processing on low capacity, on-premise facilities are numbered.

The integrated HPD/HPC technologies at NCI, supported by comprehensive data management practices, have created a new computational environment that facilitates vastly improved and enhanced processing of geophysical data. We predict that future processing of geophysical data will involve similar HPD/HPC infrastructures that will bring the users and the computation to national scale cohesive geophysical data sets that can be accessed within realistic time frames and processed at varying resolutions and at scales ranging from national to prospect. In the future, geophysical processing will routinely be multidimensional and consider a wide range of likely scenarios: uncertainty quantification and sensitivity analysis will become an inherent part of the modelling process and assist in more accurate and reliable definition of our mineral, energy, and groundwater resources. Further, where HPD/HPC infrastructures also contain relevant detailed environmental data sets, the potential impacts of any resource development can be evaluated prior to development and then monitored throughout the construction and production phases and beyond.

### **Acknowledgements**

The authors wish to acknowledge funding from the Australian Government Department of Education, through the National Collaboration Research Infrastructure Strategy (NCRIS), and Education Investment Fund (EIF) SuperScience Initiatives through the NCI, ANDS, RDSI, and NeCTAR projects. Irina Bastrakova and Carina Kemp publish with the permission of the Chief Executive Officer, Geoscience Australia.

### **Websites Cited**

All the websites were accessed January 12, 2016.

[http://www.ga.gov.au/corporate\\_data/73355/Seismic\\_and\\_magnetotelluric\\_surveys.pdf](http://www.ga.gov.au/corporate_data/73355/Seismic_and_magnetotelluric_surveys.pdf)

[http://www.ga.gov.au/metadata-gateway/metadata/record/gcat\\_69353](http://www.ga.gov.au/metadata-gateway/metadata/record/gcat_69353)

[http://www.ga.gov.au/corporate\\_data/79134/Rec2014\\_014.pdf](http://www.ga.gov.au/corporate_data/79134/Rec2014_014.pdf)

[http://www.ga.gov.au/corporate\\_data/70282/70282\\_A3.pdf](http://www.ga.gov.au/corporate_data/70282/70282_A3.pdf)

<http://vgl.auscope.org/VGL-Portal/gmap.html>

<http://nci.org.au/data-collections/data-collections/>

<http://geonetwork.nci.org.au>

<http://geonetwork.nci.org.au/geonetwork/srv/eng/metadata.show?id=14&currTab=simple>

<http://geonetworkrr2.nci.org.au>

<http://resource.geosciml.org/>

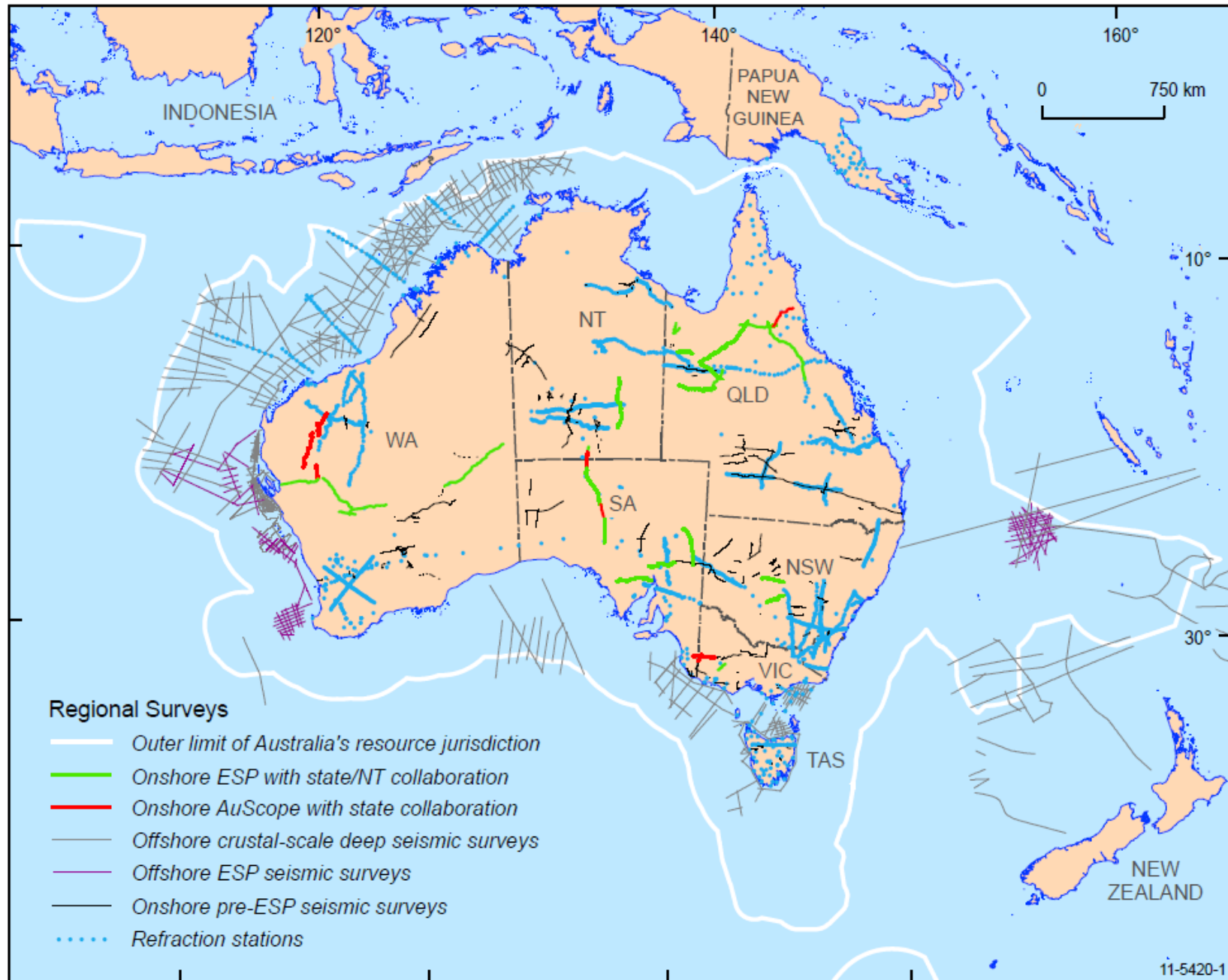
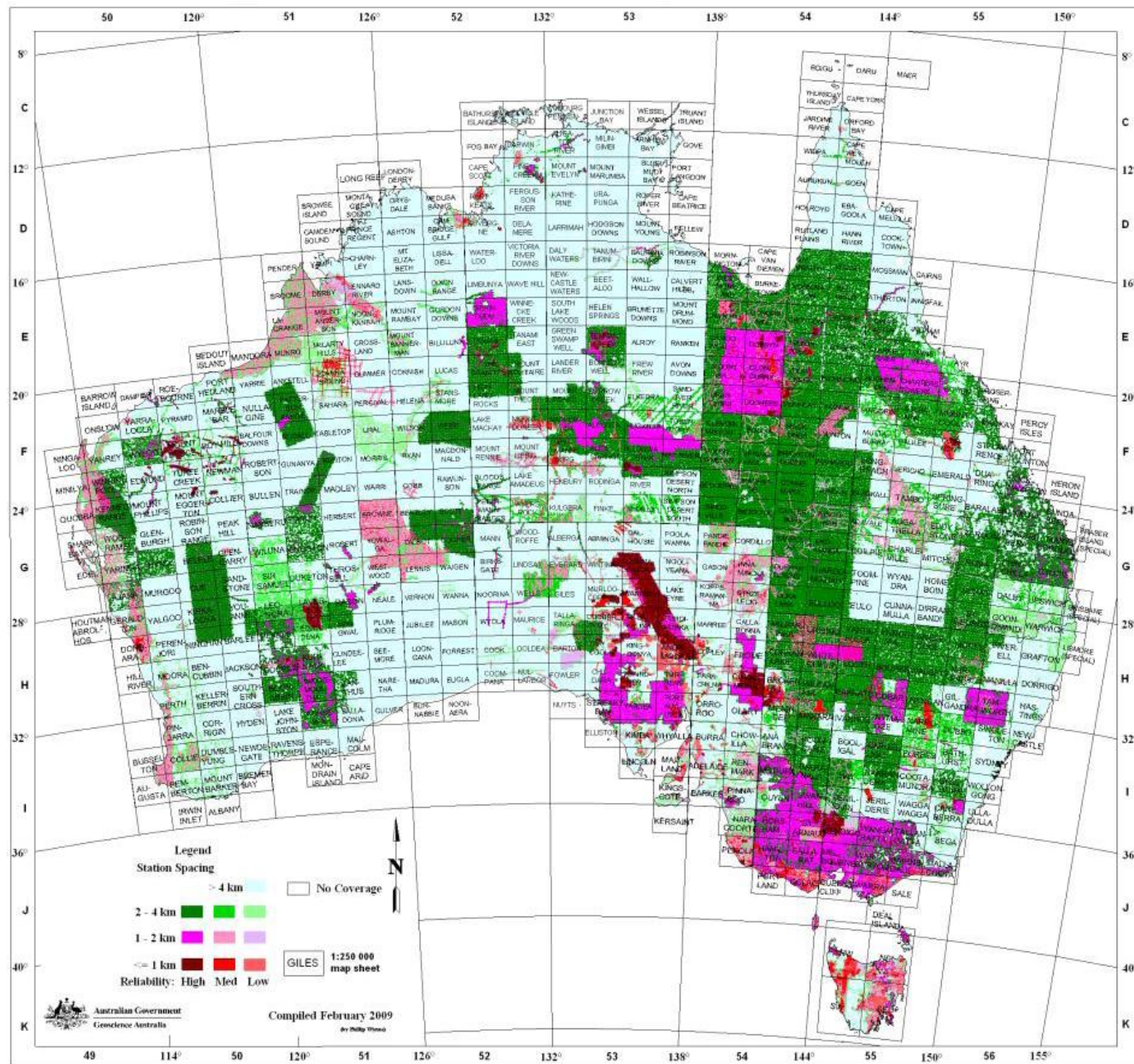


Figure 1. Map of Australia showing GA holdings of onshore and offshore deep-crustal seismic data. ESP, Energy Security Program. (Source: [http://www.ga.gov.au/corporate\\_data/73355/Seismic\\_and\\_magnetotelluric\\_surveys.pdf](http://www.ga.gov.au/corporate_data/73355/Seismic_and_magnetotelluric_surveys.pdf) )





## GAMMA-RAY DATA ACQUISITION BY GEOSCIENCE AUSTRALIA & STATES

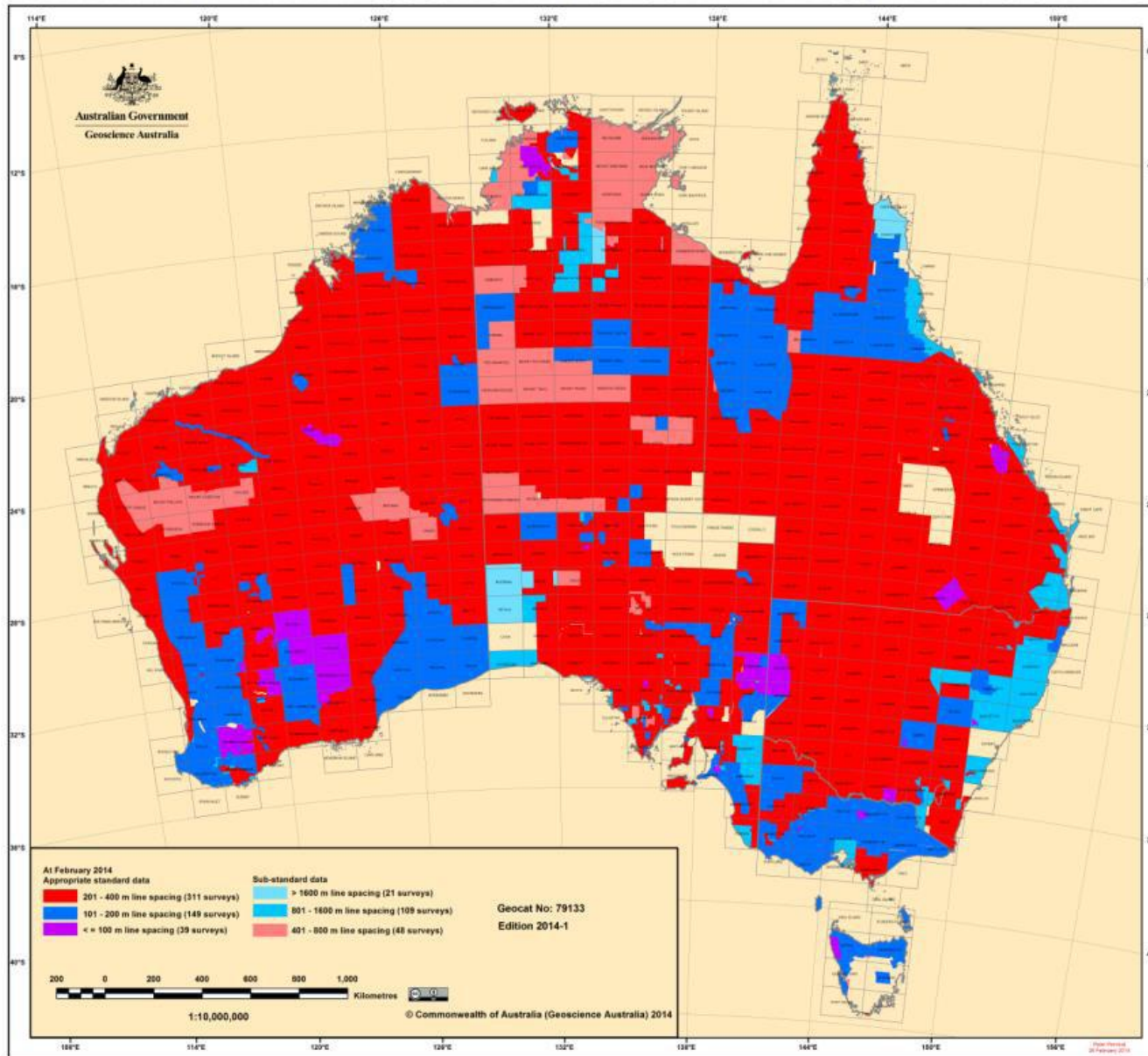


Figure 3. Airborne aeromagnetic and gamma-ray (radiometric) data acquisition index. (Source: [http://www.ga.gov.au/corporate\\_data/79134/Rec2014\\_014.pdf](http://www.ga.gov.au/corporate_data/79134/Rec2014_014.pdf) )



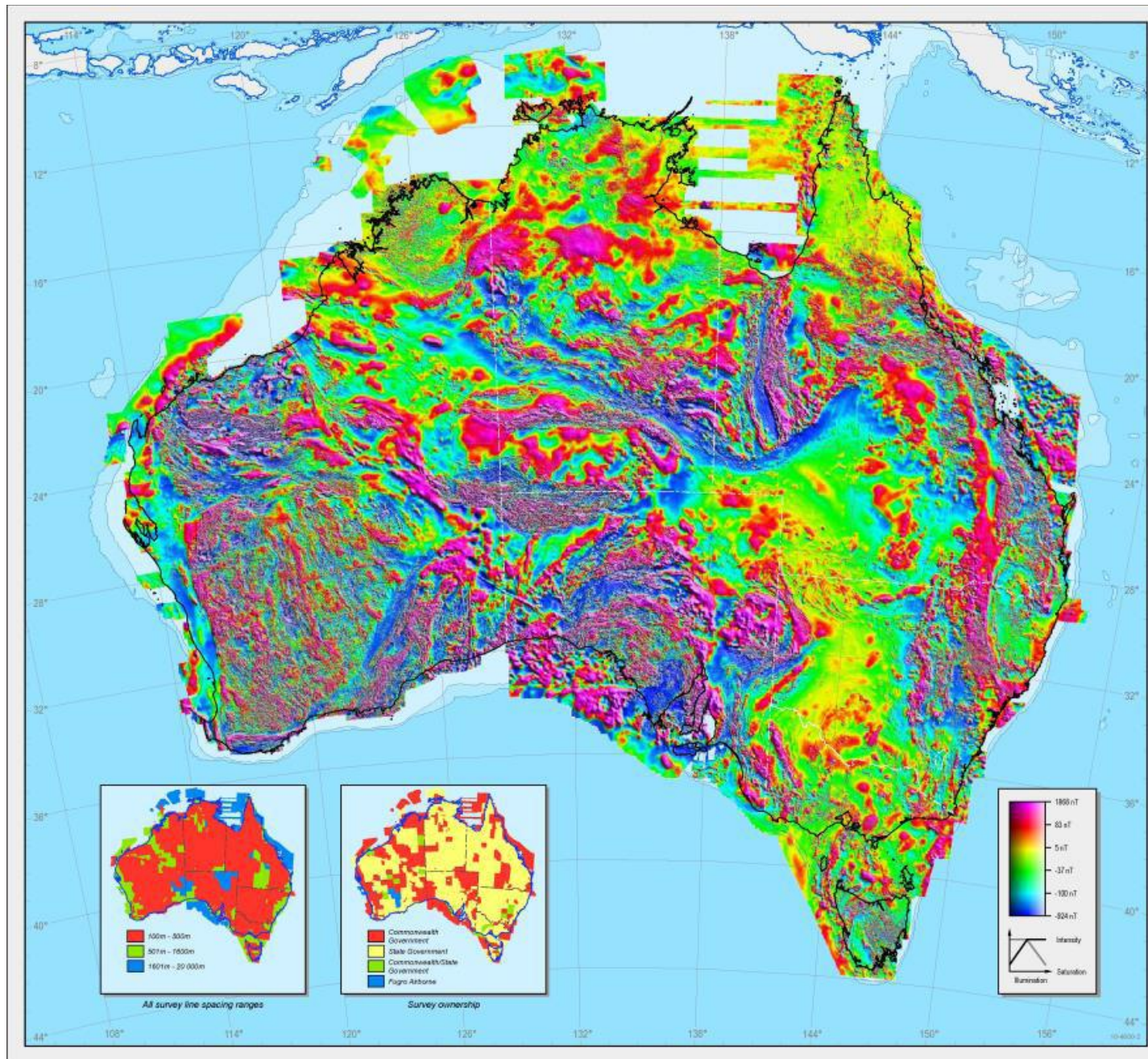


Figure 4. Magnetic Map of Australia, with line spacing and survey ownership inset. (Source: [http://www.ga.gov.au/corporate\\_data/70282/70282\\_A3.pdf](http://www.ga.gov.au/corporate_data/70282/70282_A3.pdf) )

[Single Metadata](#) | [Bulk Metadata](#)

### Metadata Publishing Form

Select your Project

**Note 1.. You have to select the project first otherwise no content will be saved.**

Insert Metadata

Metadata

Constraint

Identification

Data Quality

Maintenance

Spatial Representation

Reference System

Content Information

Distribution

Extension Information

Extent Information

File Identifier

CharacterSet

Hierarchy Level

DateTimeStamp

Metadata Standard Version

Individual Name

Position Name

Phone - Facsimile

Address - City

Address - Postal Code

Address - Email

Language

Parent Identifier

Hierarchy Level Name

Metadata Standard Name

Dataset URI

Organisation Name

Phone - Voice

Address - Delivery Point

Address - Administrative Area

Address - Country


Role

Figure 5. Screen shot of metadata creation page.

Catalog Services

dap.nci.org.au/thredds/remoteCatalogService?comm

Search



**NCI** NCI THREDDS Server  
THREDDS Data Server  
NATIONAL COMPUTATIONAL INFRASTRUCTURE  
nci.org.au

**Catalog** <http://dapds00.nci.org.au/thredds/catalog/rr2/gravityMap/catalog.xml>

**Dataset:** [gravityMap/onshore\\_only\\_Bouguer\\_geodetic\\_reprojected\\_fixed.nc](#)

- *Data size:* 81.57 Mbytes
- *Data type:* GRID
- *ID:* rr2/gravityMap/onshore\_only\_Bouguer\_geodetic\_reprojected\_fixed.nc

**Documentation:**

- [README](#)
- *rights:* [Licence](#)

**Access:**

1. **WCS:** [http://dapds00.nci.org.au/thredds/wcs/rr2/gravityMap/onshore\\_only\\_Bouguer\\_geodetic\\_reprojected\\_fixed.nc](http://dapds00.nci.org.au/thredds/wcs/rr2/gravityMap/onshore_only_Bouguer_geodetic_reprojected_fixed.nc)
2. **WMS:** [http://dapds00.nci.org.au/thredds/wms/rr2/gravityMap/onshore\\_only\\_Bouguer\\_geodetic\\_reprojected\\_fixed.nc](http://dapds00.nci.org.au/thredds/wms/rr2/gravityMap/onshore_only_Bouguer_geodetic_reprojected_fixed.nc)
3. **OPENDAP:** [http://dapds00.nci.org.au/thredds/dodsC/rr2/gravityMap/onshore\\_only\\_Bouguer\\_geodetic\\_reprojected\\_fixed.nc](http://dapds00.nci.org.au/thredds/dodsC/rr2/gravityMap/onshore_only_Bouguer_geodetic_reprojected_fixed.nc)
4. **HTTPServer:** [http://dapds00.nci.org.au/thredds/fileServer/rr2/gravityMap/onshore\\_only\\_Bouguer\\_geodetic\\_reprojected\\_fixed.nc](http://dapds00.nci.org.au/thredds/fileServer/rr2/gravityMap/onshore_only_Bouguer_geodetic_reprojected_fixed.nc)
5. **NetcdfSubset:** [http://dapds00.nci.org.au/thredds/ncss/rr2/gravityMap/onshore\\_only\\_Bouguer\\_geodetic\\_reprojected\\_fixed.nc](http://dapds00.nci.org.au/thredds/ncss/rr2/gravityMap/onshore_only_Bouguer_geodetic_reprojected_fixed.nc)

**Dates:**

- 2014-12-04T02:30:53Z (modified)

Figure 6. Geophysical data publishing interface through THREDDS.



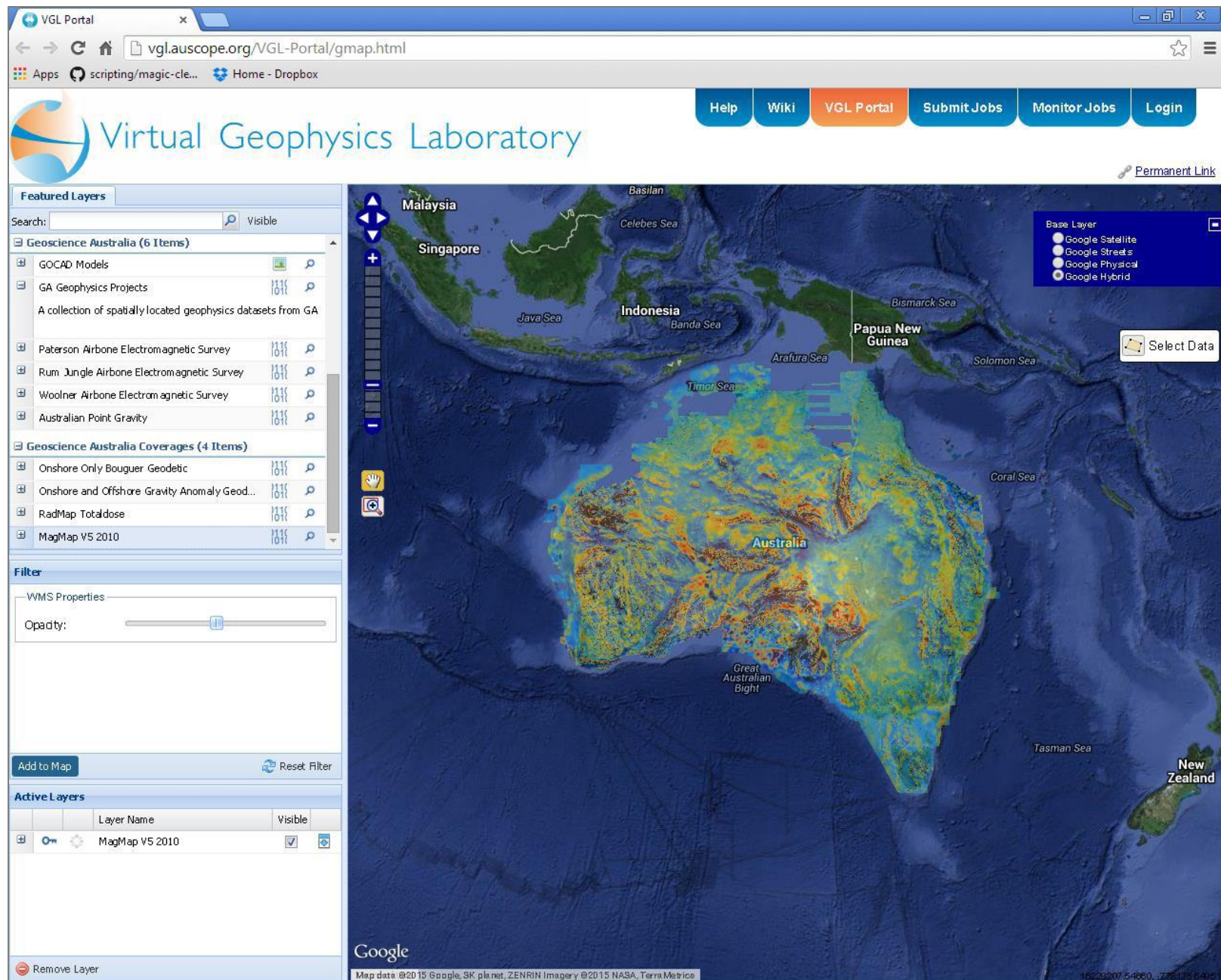


Figure 7. Screen capture of the Virtual Geophysics Laboratory. (Source: <http://vgl.auscope.org/VGL-Portal/gmap.html> )

	Current data holdings size (TB)	Planned data holdings Size (TB)	Current gridded/section data formats	Planned point/raw data formats
Onshore seismic	2	50	SEG Y	HDF5
Offshore seismic	2	200	SEG Y	HDF5
MagnetoTellurics	0	20	NetCDF	HDF5
Gravity	0.1	1	NetCDF	HDF5
Magnetic	2	10	NetCDF	HDF5
Radiometrics	2	10	NetCDF	HDF5
Digital Elevation Models	1	4	NetCDF	HDF5
Airborne Electromagnetic	0.1	5	NetCDF	HDF5
<b>TOTAL</b>	<b>9.2</b>	<b>300</b>		

Table 1. Summary of data sets in the NCI Australian Geophysical Data Collection.