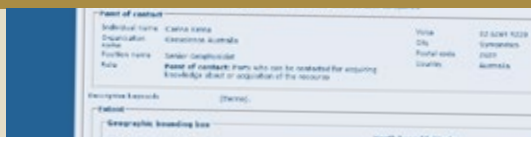
Australian Government
Geoscience Australia

AUSGEO news
ISSUE 109 Mar 2013

High Performance Computing

New Frontiers, New Paradigms, New Science



Lesley Wyborn

Reviewing Geoscience Australia's forays into High Performance Computing

In September 2010 Geoscience Australia began a trial project to determine if, and potentially how, it could evaluate the potential to utilise High Performance Computing (HPC) to improve seamless access to its large national data sets to enable processing at higher resolution within faster timeframes. The approach consisted of:

- A partnership with the National Computational Infrastructure (NCI) facility at the Australian National University to test Geoscience Australia applications in HPC environments
- An eResearch collaboration project with CSIRO to trial new tools and techniques in online data access and processing
- Several collaboration projects with the national research sector via the Department of Industry, Innovation, Science and Tertiary Education (DIISRTE) eResearch Infrastructure Programs, in particular with the Australian National Data Service (ANDS) and the National eResearch Collaboration Tools and Resources (NeCTAR) Project.

"Geoscience Australia's data and information products are being used to support decision making in the environmental and resource science domains"

Two discipline areas (Earth Observation and Onshore Geophysics) were involved in the early trials that were quickly regarded as resounding successes. Within a few months it was proven that Geoscience Australia could analyse much larger data sets, at higher resolutions and faster utilising the HPC systems at the NCI. Progressively more projects became involved and as of March 2013 there are now 20 active projects across Geoscience Australia that are utilising HPC to improve the quality and resolution of scientific outputs and hence create better business outcomes. Examples of some of these current Geoscience Australia projects utilising the NCI facility include:

- A Geophysics Inversion capability which enables high resolution geophysics inversions of magnetic, gravity, Airborne Electro Magnetic (AEM) and Magneto Telluric (MT) data
- A Geoscience Hazards capability that enables probabilistic modelling of potential impacts of earthquakes, tsunamis, storm surges and severe wind events
- A Geodetic Capability whose activities include measuring surface deformation of the Australian Continent and updating of Australia's National Coordinate reference system
- A Geothermal Capability which enables modelling and prediction of heat flow around buried granites
- An Earth Observation Capability which is creating a seamless, time series data cube of 30 years of Landsat data to enable empirical documentation of changes in land cover.

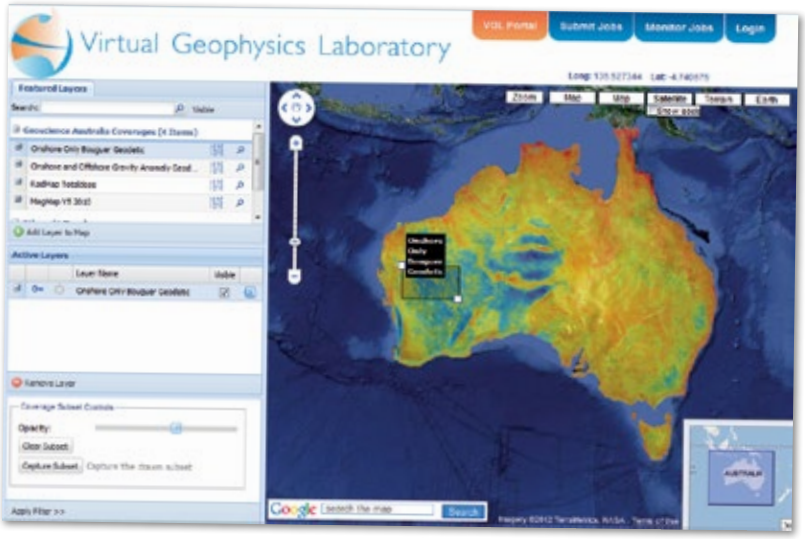


Figure 1. The user interface to the Virtual Geophysics Laboratory

Increasingly Geoscience Australia staff are realising the new opportunities that HPC facilities such as the NCI offer. An end goal is for quantitative, integrated assessment and modelling of complex earth science systems to enable us to respond to decadal challenges in Earth Systems Science including water security, sustainable development of our minerals and energy resources, community safety and emergency management, and sustainable environmental management. Effective participation in these new frontiers requires a capacity in HPC to access multidisciplinary data at high spatial and temporal resolutions and to cover large geographic extents (in 3D wherever possible). It also necessitates considering a range of likely scenarios with uncertainty quantification and sensitivity analysis becoming an inherent part of the modelling process.

A related development is that increasingly Geoscience Australia's data and information products are being used to support decision making in the environmental and resource science domains and there is an increasing demand for transparency in the decision making process with organisations required to justify why they made certain decisions or provided certain advice. As HPC enables the creation of multiple models and scenarios it becomes essential that provenance workflows are also developed to store the input data and capture what processing was done to these data during any modelling and/or processing session.

However, after two and a half years of operating at NCI, there is a realisation that although the NCI facility offers the computational capability to undertake complex analyses, the transition to the HPC environment is not easy. It is a new paradigm. It is not just a matter of transferring existing internal work practices: these have to be radically transformed. Geoscience Australia's experience, and that of many other organisations, is now showing that it is becoming increasingly difficult to scale existing earth science methods and software. The transition to HPC is creating new requirements for:

- New tools and applications that operate in massively parallel HPC environments
- Cohesive, value-added, nationally calibrated data sets that can be accessed within realistic time frames
- Creating online virtual laboratories that allow seamless access to distributed data, software and computer services and can also automatically generate provenance workflows to enable all aspects of scientific process to be automatically captured (figures 1 and 2).

This article outlines how Geoscience Australia is approaching these new requirements and then proposes some future directions that will exploit new opportunities that the new computing environments are enabling.

The requirement for a new generation of Earth Science HPC tools

In recent years, with the ending of Moore's Law, computer hardware trends are towards increased parallelisation and larger memory. However, although HPC can meet the high computational demand that is required for the integration and analysis of data at high resolution and at larger geographic extents, there are very few tools that can fully exploit it. It is becoming increasingly difficult to use traditional Geographic

Information System (GIS) tools on HPC systems and many of the current commercial processing packages in use for earth scientists are struggling to scale. These packages have lagged behind computer hardware trends and are more biased towards a reliance on serial processing and/or can only access a limited number of cores (Bryan, 2013).

Users of HPC therefore have little choice other than to develop new codes or more preferably join in global community groups that are developing Open Source software to tackle similar Geoscientific problems at the scales that meet their requirements. The use of community open source codes in the earth science domain is growing and Geoscience Australia is collaborating with some of these initiatives such as the GEM (Global Earthquake Model) tools and the TOUGH (Transport of Unsaturated Ground Water and Heat) software.

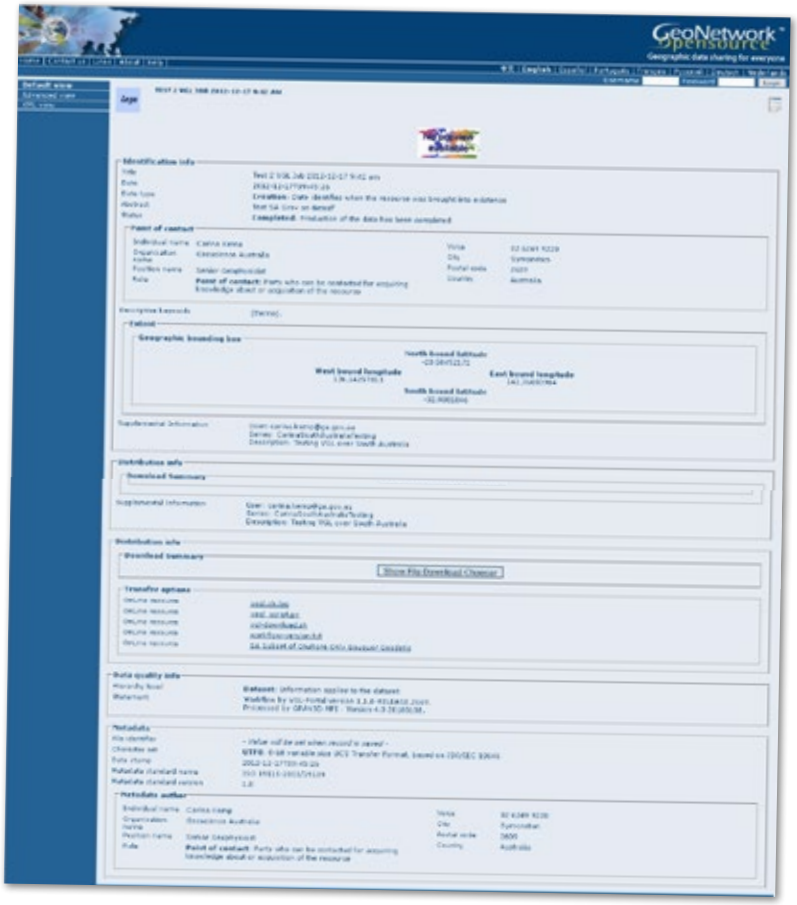


Figure 2: The provenance metadata record that is captured automatically throughout the workflow.

The level of programming skills for HPC environments required is substantially high and is very specialised. Once developed, parallel programs can also be difficult to debug and inevitably the builders of the codes become the only users. There is quite a lag time in having groups adjust to the new environment (Bryan, 2013). It is not just a matter of learning new computer language, but also new operating systems, job schedulers, etc. For Geoscience Australia, the scale of the system often means that entirely new workflows have needed to be developed. The initial transition has required specialised programmer support and there is a need for persistence and commitment to resist the temptation to go back to previous ways of doing things. To many, it has been a non-trivial exercise which requires scientists to know their code (for example, memory requirements, wall time, storage requirements), know how to access their data and to actually understand what is needed to get their model/code installed and what are the dependencies. In summary, operating in the new HPC environments requires the new generation of quantitative scientists to be more mathematically and computer literate than ever before.

The Requirement for High Performance Data Arrays/Data Cubes

It is clear that the move to HPC creates new opportunities for earth scientists to transform the way they do their science and to undertake cross-disciplinary science at much larger scales. Because of the larger capacity of HPC, it is no longer necessary for data to be averaged and subsampled: it can be analysed to its fullest resolution at national or even global scales. Much larger data volumes can be analysed in single passes and at higher resolution: large scale cross domain science is now feasible.

However, as the systems increase in capacity, access to large earth

science data is becoming a pain point. Many current issues with data access are historic and stem from the limitations of early data storage systems. As storage was so expensive, metadata was usually stored separate from the data and attached as a 'readme' file. Likewise, attributes that defined uncertainty, reliability and traceability were recorded in laboratory note books and rarely stored with the data. Many earth and environmental science datasets are fragmented across multiple institutions and the data sets themselves are notoriously heterogeneous and difficult to aggregate. Although many 'big data' collections are of the order of terabytes and even petabytes in size, they actually consist of small individual files each less than 2 gigabytes that have to be accessed individually via associated metadata records stored separately from the data.

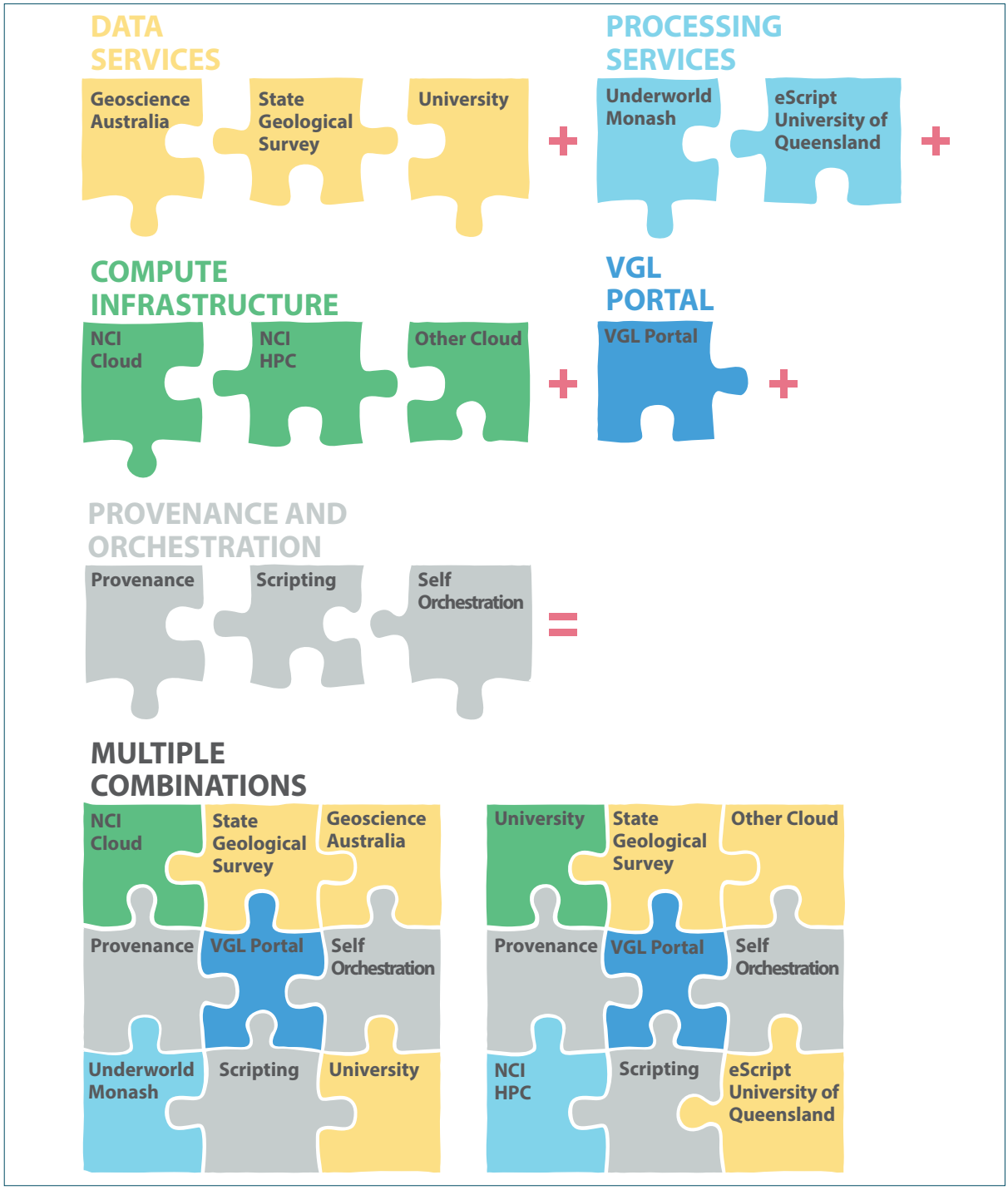


Figure 3: The components of the Virtual Geophysics Laboratory.

Thus, in an HPC environment, it is becoming increasingly difficult to dynamically access these data sets within realistic time frames and there are new requirements for data to be transformed and aggregated into High Performance Data (HPD) data sets that comprise cohesive, nationally calibrated data arrays/clouds or data cubes. Further, for data access and assimilation to be improved, data will need to be self describing. For heterogeneous data to be rapidly integrated, attributes such as reliability, uncertainty and traceability will also need to be systematically recorded with each observation. International standards for machine readability of data become critical and are the enablers of integration. It requires highly specialised skills to create these large volume standard HPD arrays and cubes, but once created, they are easier to access and use by the non-specialist.

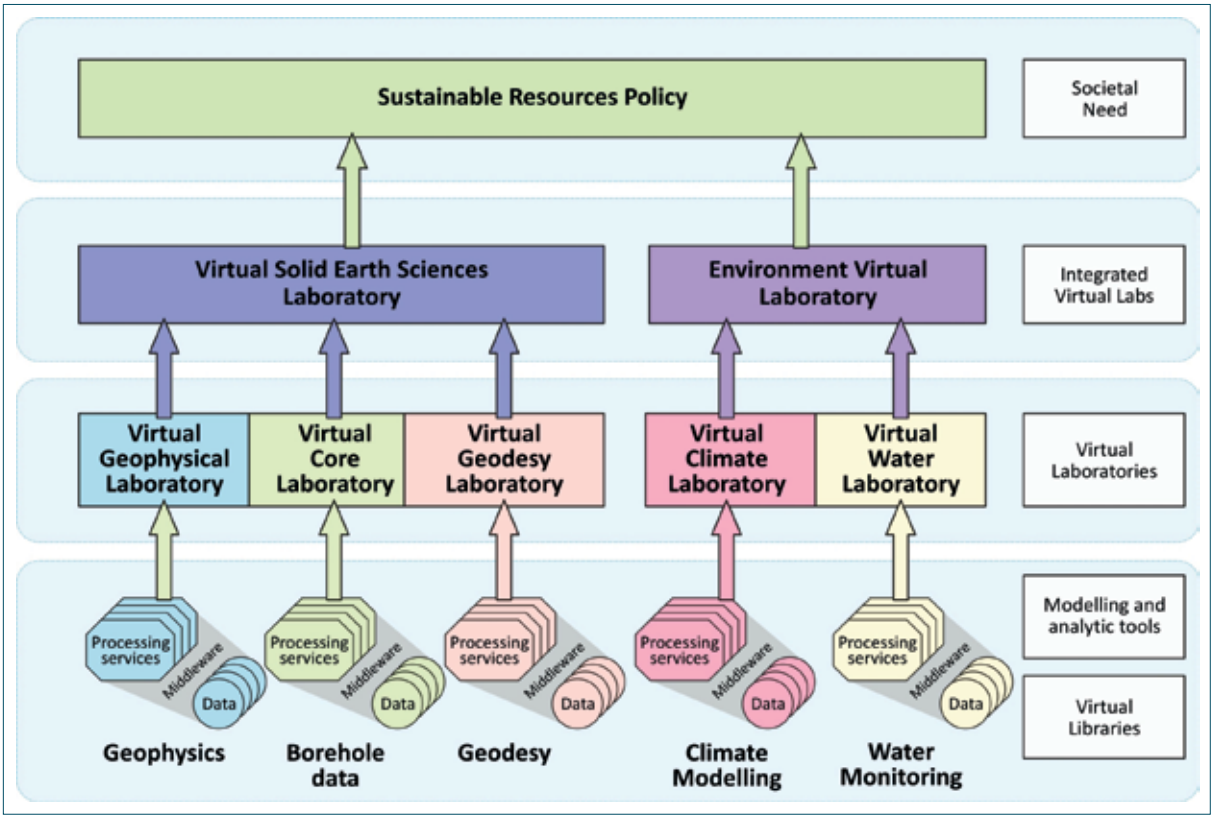


Figure 4: The plan for a national geoscience information infrastructure—linking multiple virtual laboratories via seamless workflows.

The Requirement for Virtual Laboratories to lower barriers to entry

It is clear that HPC is quite a complex environment requiring highly skilled programmers and sophisticated integrated data cubes/data arrays. Increasingly, virtual online laboratories are being seen as a means of enabling large scale online processing to take place transparently across distributed data, software and compute resources. A virtual laboratory has three phases: data selection; selecting and tuning the processing algorithms; and submitting the job to the most appropriate computational facility and monitoring.

The entire infrastructure can provide geoscientists with an integrated environment that enables seamless access to distributed data libraries and loosely couples these data to a variety of processing tools. A laboratory can link to a variety of compute resources that span from petascale HPC systems to private and commercial clouds and to local desktop. The user accesses the laboratory through an intuitive user-centred interface that enables real-time seamless linkage between all components. A provenance workflow can be automatically generated in the background by capturing

information on all inputs to the processing chain, including details about the user and their organisation.

The DIISRTE NeCTAR project is funding a series of 11 such Virtual Laboratories and one of these NeCTAR projects is a collaboration between Geoscience Australia, CSIRO, the National Computational Infrastructure, Monash University, the Australian National University, and the University of Queensland to build a Virtual Geophysics Laboratory (VGL). The VGL provides a distributed system whereby a user can enter an online virtual laboratory and seamlessly connect geoscience data, processing software, and computational hardware available from a variety of cooperating, but independent, Australian Government and university organisations (figure3).

"Globally the earth scientists are realising that HPC and new data aggregation technologies offer far more than 'bigger, faster, and at higher resolution'"

The VGL provides a web mapping interface to discover and filter the data sources using spatial and other attributes (figure 1). Once the data is selected the user is not required to download the data—the VGL collates the service query information for later in the processing workflow where it will be staged directly to the computing facilities. Throughout the workflow, provenance information is collected and captured into a metadata record where possible attributes collected are compliant with the ANZLIC profile of the ISO 19115 metadata standard (figure 2). This provenance workflow enables a team to not only keep track of jobs they have submitted, it also enables spatial displays of who ran what, where and when. The system is quite flexible and is glued together via python scripts that can either be reused as is, or adapted by other users.

VGL is therefore more an infrastructure rather than application. It orchestrates the linking of multiple data sets that are available as Open Geospatial Consortium (OGC) compliant web data services to a variety of software resources that are also available as services (figure 3). Thus VGL acts as a broker and can seamlessly integrate data services from distributed locations, meaning that less time is spent locating, downloading and reformatting disparate data sets into integrated standardised data sets that can then be accessed by HPC.

The benefit of the provenance workflow is that all products produced are transparent. As the metadata record captures all input files, including any changes to processing algorithms, time of extraction of the data from the databases, who ran what and where, the components of workflow can be reused by others wanting to run similar workflows. The greatest benefit is that all procedures used are accessible, verifiable and can be used by other investigators to test the results.

Where to next: a National Geoscience Information Infrastructure

Geoscience Australia's experience in successfully trialling HPC to enable new science and new business outcomes, particularly in analysing large volume data sets is not alone. Globally earth scientists are realising that HPC and new data aggregation technologies offers far more than 'bigger, faster, and at higher resolution'. The new paradigm is an opportunity for a transformative change in the way in which geoscience data analysis is routinely conducted and that it has the capacity to create new ways of doing collaborative, 'transparent' analysis in which all relevant data from distributed data sources can be harnessed online.

Infrastructures, such as virtual laboratories, can lower the skills barriers to entry for HPC and can increase uptake. As is illustrated in figure 4, the VGL is a prototype geoscience infrastructure that enables geophysics data to be linked transparently to a variety of software and compute resources. There is no reason why similar systems could not be built for geodesy, boreholes, climate and water and already there a plans to build some of these. Integrated and linked through the use of OGC/ISO standards, they can become a National Geoscience Information Infrastructure and capable of integrating with other major infrastructures including the National Plan for Environmental Information, the ANZLIC One ANZ Foundation Spatial Data Framework and the eMarine Information Infrastructure.

References

Bryan, B. 2012. High-performance computing tools for the integrated assessment and modelling of social-ecological systems. Environmental Modelling and Software, 39, 295-303.

For more information

email ausgeomail@ga.gov.au

Realated articles and websites

Virtual geophysics Laboratory
http://vgl.auscope.org/VGL-Portal/gmap.html

National Computational Infrastructure
http://nci.org.au/

The National Plan for Environmental Information
www.environment.gov.au/npei/index.html

ANZLIC One ANZ Foundation Spatial Data Framework
http://spatial.gov.au/system/files/public/resources/anzlic/ANZ_FoundationSpatialDataFramework_(FinalWeb).pdf

eMarine Information Infrastructure
http://imos.org.au/emii.html