Geophysical Research Abstracts
Vol. 14, EGU2012-7045, 2012
EGU General Assembly 2012

Why the Petascale era will drive improvements in the management of the full lifecycle of earth science data.

L. Wyborn
Geoscience Australia, Canberra, Australia (lesley.wyborn@ga.gov.au)

The advent of the petascale era, in both storage and compute facilities, will offer new opportunities for earth scientists to transform the way they do their science and to undertake cross-disciplinary science at a global scale. No longer will data have 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, in general, earth sciences have been slow to capitalise on the potential of these new petascale compute facilities: many struggle to even use terascale facilities. Our chances of using these new facilities will require a vast improvement in the management of the full life cycle of data: in reality it will need to be transformed.

Many of our current issues with earth science data 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 recoded in lab note books and rarely stored with the data. Data were routinely transferred as files.

The new opportunities require that the traditional discover, display and locally download and process paradigm is too limited. 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 need to be systematically recorded with each observation. The petascale era also requires that individual data files be transformed and aggregated into calibrated data arrays or data cubes. Standards become critical and are the enablers of integration.

These changes are common to almost every science discipline. What makes earth sciences unique is that many domains record time series data, particularly in the environmental geosciences areas (weathering, soil changes, climate change). The data life cycle will be measured in decades and centuries, not years. Preservation over such time spans is quite a challenge to the earth sciences as data will have to be managed over many evolutions of software and hardware. The focus has to be on managing the data and not the media.

Currently storage is not an issue, but it is predicted that data volumes will soon exceed the effective storage media than can be physically manufactured. This means that organisations will have to think about disposal and destruction of data. For earth sciences, this will be a particularly sensitive issue.

Petascale computing offers many new opportunities to the earth sciences and by 2020 exascale computers will be a reality. To fully realise these opportunities the earth sciences needs to actively and systematically rethink what the ramifications of these new systems will have on current practices for data storage, discovery, access and assimilation.