A National Environmental Satellite Data Virtual Laboratory

Edward King[1], Ben Evans[2], Lesley Wyborn[3], Wenjun Wu[3], Leo Lymburner[3], Medhavy Thankappan[3], Peter Tan[3], Fei Zhang[3], Mark Gray[4], Joseph Antony[2], Muhammad Atif[2], Matt Paget[1], Stefan Maier[5], Thomas Schroeder[1]

[1]CSIRO, Canberra & Brisbane, Australia, Edward.King@csiro.au

[2]National Computational Infrastructure, Australian National University, Canberra, Australia, Ben.Evans@anu.edu.au

[3]Geoscience Australia, Canberra, Australia, Lesley.Wyborn@ga.gov.au

[4]Curtin University of Technology, Perth, Australia, [5]Charles Darwin University, Darwin, Australia

INTRODUCTION

Earth Observing (EO) sensors carried on space-borne platforms produce large (multiple TB/year) data sets serving multiple research and application communities with complementary and overlapping needs (Fig. 1). End-user groups often use similar processing steps, but differences in temporal and spatial coverage mean that there is frequently a lack of overlap between these groups. In practice this results in fragmentation of data storage and duplication of processing systems and user analysis environments. Moreover, where overlaps do exist, they can be difficult to exploit because of specific implementation differences such as agency network firewalls, incompatible storage formats and degree of intermediate processing. This problem is common across a number of existing satellite sensors and will only get worse as new sensors are launched in the future.

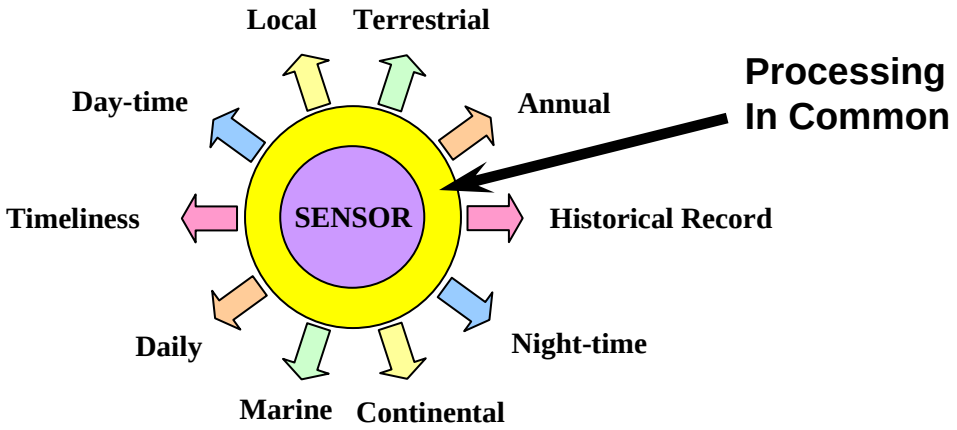

Figure 1: Competing focus areas and tensions between application communities for EO data from a single sensor type, with a layer of common data processing before application-specific product generation.

A virtual laboratory in which many data products can be constructed on-demand and from a comprehensive and high quality base data archive using common software tools is a way to counter this trend. The economies arising from sharing tools and data management can quickly overcome the cost of operating a peak facility with the capacity to support large data intensive research. By catering to a wide range of user communities a virtual laboratory also helps diminish barriers to trans-domain research, and can be an important enabler for activities such as Earth System Science that draw on multiple disciplines.

NATIONAL SATELLITE DATA SETS

Two heavily used data sets include the Moderate Resolution Imaging Spectrometer (MODIS) and LANDSAT. Two MODIS sensors were launched in 1999 and 2002 and continue to provide both day-time and night-time coverage of the globe every day in 36 spectral bands at spatial resolutions between 250m and 1km. The availability in Australia annually of 3.4TB of raw but high quality MODIS data providing observations applicable to a wide range of domains has resulted in exactly the type of fragmentation of data archives and processing systems described above. The LANDSAT sensor series began in 1972 and has provided high resolution (30m-80m) global imagery every 16 days, resulting in a raw data archive that is today approaching 200TB. When processed to several levels of data products this becomes a 1+PB archive. The large data volumes means that although a single agency (Geoscience Australia) acquires and stores the raw data, user communities tend to focus only on either small regions comprising a few individual scenes or larger regions at longer time intervals. Thus the product collection is fragmented and incomplete, and use of a particular image scene by multiple researchers can often lead to duplication in data processing and storage.

There are several other sensors of general interest to the Australian environmental research community. The Advanced Very High Resolution Radiometer (AVHRR) sensor series has been producing data since 1981 and provides the longest daily record of the Earth's surface. The first Visible Infrared Imager Radiometer Suite (VIIRS) sensor, due for launch in early 2012, is the follow-on instrument for both the AVHRR and MODIS sensors and will provide a similar data stream

to that of MODIS over the next decade. In addition there are sensors due for launch over the next decade that will produce new types of data of interest in a wide range of research areas. Preparing to manage these at a National scale is a key enabling strategy to ensure effective and efficient utilisation by the research communities.

DEMONSTRATION SYSTEMS

Through the NCI, we are establishing a national archive of MODIS and LANDSAT for the Australian region to support the base processing workflows and specialised analysis software used in advanced research. This comprehensive archive is complemented by NCI's capability to make data available in a high-performance environment. The high-speed data access is critical for both the processing and user access; enabled both within the facility on the NCI peak system and data-intensive cloud and utilizing recent improvements in connectivity to the major national and international agencies. The fast data access is a key enabler in matching data storage, processing, data-intensive analysis and data sharing using standard protocols. The excellent connectivity of the NCI to the national research backbone supports both import of the base archives (historical and in real time) and dissemination of the data products.

The virtual laboratory (Fig. 2) has been developed by the partner organisations and delivers on a number of key infrastructure goals: for the organisations operational requirements; for national infrastructure projects such as IMOS and TERN to deliver standard marine and terrestrial products; and for key new national infrastructure developments supported through the Australian Space Research Program and in demonstrating pilot production systems for research developments, such as new gridded evapotranspiration and precipitation products from CSIRO for the Bureau of Meteorology. The virtual laboratory is being set up to use not only specialised satellite processing software, but also to exploit a number of eResearch initiatives that have been developed under NeAT and elsewhere, including the RS-YABI workflow engine and the MACDDAP data aggregator. The NCI is also providing value-adding data services, such as THREDDS data servers, web interfaces and archive capabilities, to improve access and use of these data.

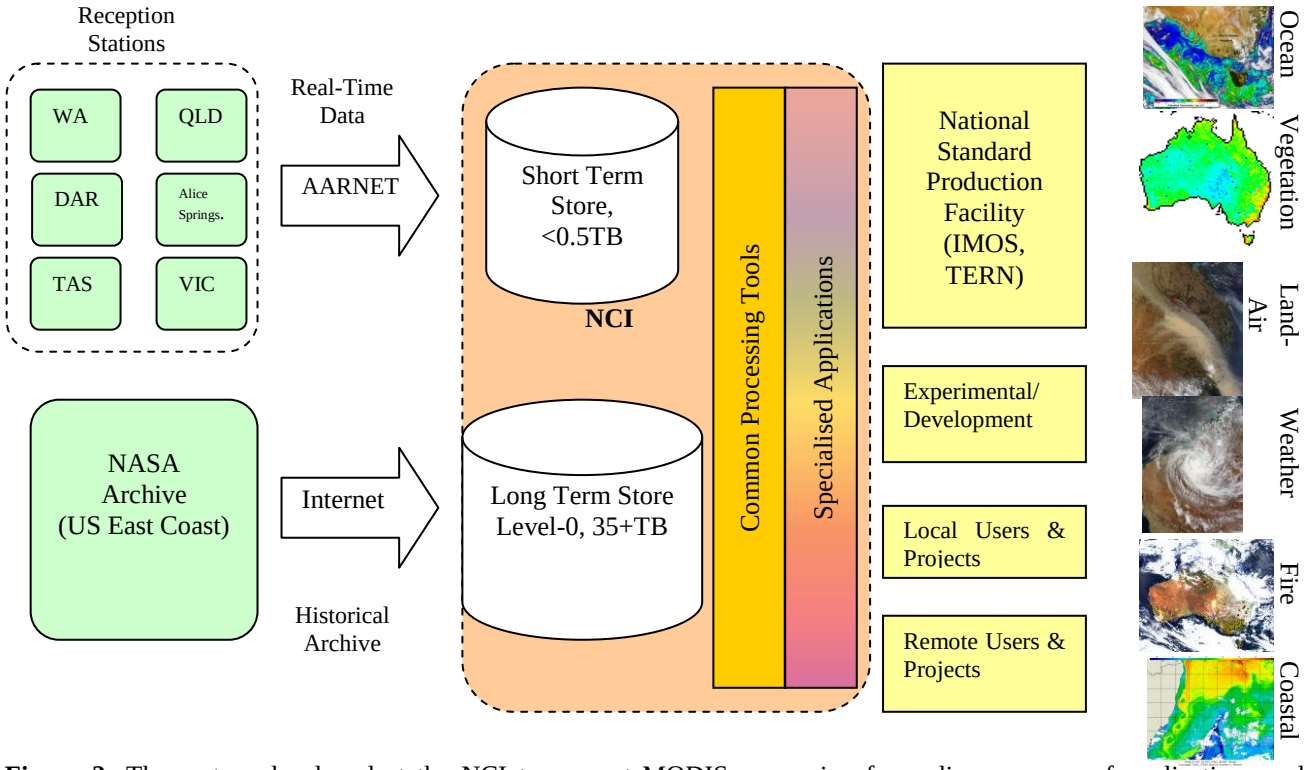

Figure 2: The system developed at the NCI to support MODIS processing for a diverse range of applications and research communities. A similar architecture is being developed for other sensors.

FUTURE WORK AND CONCLUSION

We aim to increase the range of data products produced from the existing data archives. We will also begin to support the AVHRR sensor series and prepare, from the outset, for a national approach to data management for the new VIIRS sensor when it is launched in late 2011. An important goal will be to expose the data and tools to researchers, both directly and encapsulated via advanced eResearch methods like workflow systems and domain-specific interfaces. This virtual laboratory is a means by which communities can work together to collectively overcome practical problems in common and focus on their specific research interests. It has been constructed around both computing power and data-intensive cloud facility at the NCI. The positioning of the data next to a specialised data-intensive facility is opening new potential for data analysis and opportunities for inter-disciplinary research. It is a scalable platform for collaboration serving a wide research community. This development facilitates a long-standing goal of the remote sensing community; to efficiently convert EO data into information at spatial and temporal scales that are relevant to decision makers.

BODY OF PAPER ITSELF MUST NOT EXTEND ONTO THIS PAGE

ABOUT THE AUTHORS (150-200 WORDS)

Edward King heads the Satellite Remote Sensing Facility in IMOS and leads projects in the Water for a Healthy Country and Wealth from Oceans Flagships in CSIRO. Ben Evans is the Head of the ANU Supercomputer Facility at the Australian National University. He leads projects in HPC and Data-Intensive analysis, working with the partners of NCI and the research sector. Lesley Wyborn is a Senior Geoscience Advisor at Geoscience Australia and is a member of the Australian Academy of Science National Data in Science Committee, and the Executive Committee of the Earth and Space Science Informatics Focus Group of the American Geophysical Union. Wenjun Wu, Leo Lymburner, Medhavy Thankappan, Peter Tan and Fei Zhang are members of the National Earth Observation group at Geoscience Australia. Mark Gray is a specialist in MODIS data processing and use. Joseph Antony and Muhammad Atif are data-intensive computing specialists at the NCI. Matt Paget is the TERN/AusCover data and systems coordinator. Stefan Maier is an expert in the use of MODIS data for fire and fire scar detection. Thomas Schroeder is a CSIRO research scientist developing remote sensing algorithms for inland, coastal, marine and coral reef ecosystems.

SHORT ABSTRACT (max 300 words)

A NATIONAL ENVIRONMENTAL SATELLITE DATA VIRTUAL LABORATORY

We have constructed an environment in which different research communities using large remote sensing data sets can coalesce, based on a common platform for data, workflows, and analysis tools in a high-performance environment. Earth Observing (EO) sensors carried on space-borne platforms produce large (multiple TB/year) data sets serving multiple research and application communities. The limited overlap between these end-user groups, together with the data management challenges, often leads to fragmentation of data storage and duplication of processing systems and user analysis environments. Moreover, where overlaps exist, they are often difficult to exploit because of specific implementation differences such as agency network firewalls, incompatible storage formats and the degree of intermediate processing. This problem is common across a number of existing satellite sensors and will only get worse as new sensors are launched in the future. A virtual laboratory is a means by which communities can work together to collectively overcome the problems in common and focus on their specific research interests. This virtual laboratory has been constructed around both the computing power and data-intensive cloud facility at the NCI with the support of both the IMOS and TERN NCRIS capabilities. The result is a scalable platform for collaboration in this data-rich area with far reaching interests in the research community. This development facilitates a long term goal of the remote sensing community; to convert earth observation data into information at the spatial and temporal scales that are relevant to decision makers.