

## **Workflow Services to enable a Large-Scale Temporal-Spatial Ecosystem Digital Information Service**

Submitted by:

Professor Mervyn Lynch<sup>1</sup>, Dr Peter Fearn<sup>1</sup> and Dr Edward King<sup>2</sup>

1) Department of Imaging and Applied Physics  
Curtin University of Technology  
PO Box U1987  
Perth WA 6845  
Contact Email: [m.lynch@curtin.edu.au]  
Tel: 08-9266-7540 / 7192

2) CSIRO Marine & Atmospheric Research  
GPO Box 3023  
Canberra, ACT, 2601  
Contact email: Edward.king@csiro.au  
Tel: 02-6246-5894

### **Overview** *Summarise the context that leads to this project.*

Remote sensing platforms, particularly the earth orbiting sensors of recent times, have generated a host of new data products (some 70 to 80 currently) that relate to the atmosphere, land and oceans (e.g. vegetation canopy temperature, land surface bidirectional reflectance, vegetation indices). There is a growing recognition of the utility of remotely-sensed time series for understanding biophysical and geochemical dynamics in the earth system. Model-data fusion techniques and data assimilation methods, originally developed in the weather forecasting domains, are now being taken up in the marine and terrestrial ecosystem research communities and are driving demand for enormous volumes of data. As producers of these remote sensing services, we have become acutely aware of the interest, the needs and the obstacles to the productive use of such information.

The AusCover component of the NCRIS Terrestrial Ecosystems Research Network (TERN) capability is focussed on organising remote sensing data sources and products for terrestrial ecosystems research. AusCover will enable, for the first time in Australia, the storage of these data sets online in a form that makes them directly accessible to the user community.

AusCover's role relates specifically to delivering a range of remote sensing data products that are quality assured. Its mandate does not include provision of tools and services for researchers to easily analyse and process the data sets, which may require large amounts of computation on very large amounts of data, particularly for large-scale spatio-temporal analysis. This means there are many situations where there is a significant gap between what AusCover can deliver and what many users seek.

This project aims to fill this gap by providing easy-to-use workflow tools and services that enable researchers to process AusCover data sets using the ARCS grid (or cloud) computing infrastructure. The same workflow tools will also assist AusCover data providers, allowing them to more easily process raw satellite data in order to generate derived data products in the standard data formats that users require.

This NeAT project proposes to deploy workflow systems and tools which will:

- Provide a standard Workflow system that will enable AusCover data providers and users to easily specify and run custom data processing workflows that can access data from AusCover (and other data sources) and process it on remote computing resources, including the ARCS Grid infrastructure.
- Enable end users to easily utilise the massive computing infrastructure associated with the TERN/AusCover data management project, including tools for converting data to standard formats, running standard data processing applications on remote sensing data, and

integrating TERN products with other spatial / temporal data sets (from outside the TERN project).

- Enable AusCover data providers to more easily process raw satellite data to generate standard derived data products, as well as extracting standard metadata to enable improved search capability.
- Demonstrate the functionality of this generic approach with some exemplar workflows which will involve chained processing services across large numbers of data files, e.g. time-series of large spatial data arrays, and be used to generate data products of use to many researchers.

The capacity to perform the significant computing required to generate these products directly will be a boon to research activities that seek to develop and refine application products, and will be essential to any user application that attempts to make any type of near real-time use of such products for forecasting or monitoring purposes. The processing of these types of data is inherently parallelisable as each data granule is generally independent of all others in both space and time and so processing multiple granules distributes across multiple processors very naturally. This is well suited to processing on the ARCS Grid infrastructure and its future evolution.

We are proposing to develop a generic workflow system with a suite of standard workflows that can be easily adapted to a range of user interests and applications. For example, the AusCover data sets utilised and the ancillary data sets to be acquired or accessed from a number of sources may change, but the structure of the Workflow will require minimum adjustment, since different components can be easily incorporated into the existing workflows in order to access alternative data sources and to accommodate other data formats.

**Users** *Identify the research communities and resource providers that this project serves; and the potential number of users. This should include some NCRIS capabilities or other data federating or collaborating research groups, and any institutions that will participate through setting requirements for or steering this project.*

The workflow tools and services for data processing that will be developed by this project will be utilized by both the providers of remote sensing data and associated derived data products, and the users of these data sets.

The target data providers are the partners in the following NCRIS capability areas:

- TERN AusCover, including Bureau of Meteorology, CSIRO, Geoscience Australia, Curtin University, Landgate (WA), University of Queensland, Queensland Dept. Natural Resources, RMIT University, Charles Darwin University, together with any other organisations that take the opportunity to serve data via the AusCover network.
- IMOS Australian Oceans Distributed Active Archive Center (AO-DAAC), including CSIRO Marine & Atmospheric Research, CSIRO Land and Water, Curtin University and University of Queensland. AusCover will build upon the initial work undertaken in that project and the types of processing conceived in this project are generally applicable to remote sensing data, regardless of whether it is terrestrial, marine or atmospheric.

Because this project is synergistic with Ausover and IMOS/AO-DAAC, it can act as a fulcrum in the remote sensing data domain for the activities of NCRIS/PfC/Australian National Data Service (ANDS) by providing an opportunity to influence the choice of data standards that will be supported by a major processing conduit through which much data will flow. It is proposed that AusCover will utilise the data services provided by the ARCS and the Members of ARCS (MARCS). This project seeks to exploit the opportunity to utilise the ARCS-supported National Grid to process these collocated data.

User communities that will directly benefit from these data products, and the workflow tools for custom data processing on the grid that will be developed in this project, include:

- Members of the ARC Network for Earth System Science (ARCNESS), which links individuals and groups within Australia to explore the interaction between the oceans, atmosphere, biosphere and cryosphere. The network is a partnership of over 30 Australian research organisations, with around 300 registered participants from Australia and more than 60 from overseas institutions. As an example, ARCNESS members are seeking nationally gridded products for initializing the numerical climate forecast model ACCESS.
- Other users of the AusCover data system who are developing new products or remote sensing applications. Since AusCover has not yet officially commenced, it is not possible at this stage to be specific about the number of users, though it is potentially a substantial fraction of the large community of terrestrial ecosystem researchers.
- With the growing interest in the national vegetation estate (e.g. the Greenhouse Office), the impact of climate on vegetation and the associated surface forcing (Bureau of Meteorology), and carbon sequestration, we anticipate that remote sensing will offer new information services to the research community. These agencies have indicated a strong interest in being involved and we are aware of similar needs from state agencies. We expect these user groups to be early adopters of the proposed services.
- Users of the NCRIS/IMOS/AO-DAAC who access, process and use similar remotely-sensed spatial time-series data.

**Needs** *Describe the needs of the research communities or resource providers that this project seeks to address.*

As the Earth System responds to changes in climatic forcing, time series of physical measurements are one of the primary means of monitoring and detecting changes, and the longer the time series, the more sensitive is the test. The large remote sensing data sets that are used to produce time series of physical and biological parameters at continental scale present particular processing challenges relating primarily to data volume and the need to assemble processing chains to apply a series of standardised processing steps. These challenges are an obstacle that must be overcome to begin to fully exploit the wealth of information that is recorded within these existing data sets, and the more voluminous ones to come.

TERN is charged with delivery of quality-assured products that are to be made easily accessible in common data formats and with standard metadata. For AusCover, providing data and metadata in this form requires a significant amount of processing of the very large amounts of existing base remote sensing data, as well as the new data that is constantly being produced. Providing a simple way of specifying and automating the required data processing workflows will be of significant benefit to the AusCover data providers and enable them to more easily provide a wide range of derived data products to their users. While the data provision elements of AusCover underpin this project, and are a key element required to improve data utilisation, for researchers to make best use of the data provided by AusCover, they will need easy-to-use tools for customised, server-side data processing, that can support workflows of multiple chained data processing tasks. A system that enables researchers to customise processing workflows and then distribute the processing across a grid will go a long way to unlocking the potential of the types of data that will be provided by AusCover.

It is very clear that frequently current levels of activity in ecosystems research and data collection (i.e. field programs) are seriously undersampling the systems that they are studying (e.g. health and primary production of the terrestrial ecosystem; and change detection in the terrestrial ecosystem). There is a need for more comprehensive information, more temporally sampled information, quantitative rather than qualitative data, validated data and of course impact assessment. This requires easy-to-use workflow tools that enable researchers to easily specify customized data processing

workflows that can integrate the data collected from field programs with relevant remote sensing data provided by AusCover.

Enabling the workflow system to seamlessly utilize a variety of compute resources, in particular those available through the ARCS Grid (and its planned successor, the ARCS Compute Cloud), will enable more timely data processing without the need for AusCover or its users to purchase very large, dedicated compute resources to meet their data processing needs. Since the data provided by AusCover will be distributed across multiple sites, this also enables processing to be done on compute resources that are co-located with the data where possible, allowing faster and more efficient data access. By building a data processing facility that is designed, from the outset, to work with both AusCover and IMOS/AO-DAAC provided distributed data, there is more chance that it will be sufficiently generic to meet other similar needs in future.

A system that enables researchers to customise processing workflows and then distribute the processing across a grid will go a long way to unlocking the potential of the types of data that will be provided by AusCover.

The following are a few examples of research applications that involve large-scale data processing to produce a one-off data set for identified end users, and then progressively refining it and keeping it up to date.

1. Re-processing of the MODIS satellite archive (25 TB) to improve fire hotspot and fire scar histories for the past decade. This sort of task relies on a degree of statistical validation and algorithm tuning, so the capacity to run and rerun progressively refined algorithms is very much an experimental research process. (Charles Darwin University, Landgate, Bushfire CRC, Commonwealth Department of Climate Change)
2. Another task relying on progressive refinement of algorithms run on large data sets is processing of the NOAA HRPT archive (10 TB) to produce a 25-year history of vegetation green-ness, temperature dynamics, and cloud cover. (CSIRO Marine and Atmospheric Research, Bureau of Meteorology, state and federal government agencies)
3. Model data assimilation for surface energy balance modelling to improve water resource assessment and understanding of biogeochemical (*ie.* carbon) fluxes. This is a task which is both computationally intensive, through the extensive modelling, and data intensive due to the diverse range of data inputs. (CSIRO Water for a Healthy Country & Climate Adaptation Flagships, together with Universities of Queensland, NSW and Melbourne, Bureau of Meteorology Water Division, Dept. of Climate Change)
4. Integration of satellite products (from AusCover) with *in situ* observations from field sites, to support a range of analyses including improved temporal management, denser sampling of the range of ecosystems, demarcation of ecosystem boundaries, vegetation health (as measured by biomass and its temporal variability) with seasonal and interannual climate variability. (WA Department of Environment and Conservation, Curtin University, and many other government agencies and users of TERN data).
5. Analysis of dust storms over the northern agricultural region of WA and the wider pastoral rangelands of central Australia, including source location, storm track, destination and estimation of top soil load transported in the storm (DAFWA, Curtin University)
6. Study of the impact of bushfire smoke on people with respiratory diseases, using bushfire plume trajectory data and a set of statistics extracted from imagery (*eg.* estimated exposure, frequency) merged with patient data (WA Dept of Health, Curtin University)

The proposed workflow services will make it much easier to handle the large-scale data processing required by these, and many other, applications.

**Services** *Describe the result of the project in terms of the service(s) that will be implemented and demonstrated by the project and which could be operated in an ongoing fashion; and the proposed operator of each service.*

A generic Workflow system will be deployed for processing of AusCover data, which will enable the applications in the data processing workflow to be executed on the ARCS Grid (and its proposed evolution, the ARCS Compute Cloud), and will support data access from distributed data servers, including the ARCS Data Fabric and OpeNDAP servers of the type already deployed in the IMOS/AO-DAAC and NeAT/MACDAP projects.

There are many Workflow tools that could be used to implement such a service. ARCS (with advice from NeAT and others) will work with the other participants in this project to select an appropriate workflow tool and ensure that it becomes an ARCS-supported, production service that can utilize ARCS compute and data infrastructure, and can be used as the basis for the services developed in this project.

This is an area where the proposed NeAT project will significantly enhance and extend the utility of AusCover activities and provide a system that will be of broad utility. It will also act as an exemplar to other research communities for the automation of data processing workflows, and for the use of the ARCS Grid resources as the computing and communications infrastructure for processing such workflows. Similarly it is a project which will demonstrate the value of the TERN/AusCover distributed data provision over the network.

The project will package and deploy a suite of standard spatial processing tools that will be made accessible from the workflow system. These will provide a mixture of sensor-specific functionality (e.g. calibration and geolocation) and generic spatial operations such as reprojection, remapping, sub-setting, sub-sampling and aggregation (*ie.* spatial averaging over regular grids or irregular regions of interest). The unifying feature of these services would be the common data models and metadata standards that would have to be adopted in order to cascade these services together to establish spatial processing workflows. A significant amount of effort would be required to ensure this uniformity in the processing chain, *eg.* compatible spatial gridding, data formats and metadata. Deploying standardised versions of these services in a Grid environment so that they can be reused by different applications would provide a huge improvement in both processing efficiency and application development time.

A set of predefined workflow instances will be developed using this suite of tools, that will enable AusCover data providers to generate standard derived data products from raw remote sensing data, that will be useful to the broader earth system science community. At their most generic, these could typically include various types of statistical evaluations over user-specified regions of interest, or time-series analyses. More sensor specific examples might be implementing standard base processing chains for specific sensors, such as MODIS L0 to L1B, so that large volumes of data granules can be processed consistently for analysis.

The project will also provide a mechanism for users to easily create customised processing workflows using the workflow system (such as those specified in the Needs section), and to have these workflows executed on the ARCS Grid. The project will work with some end user groups to develop some exemplar workflows of this kind.

It is envisaged that these earth system science tools will be operated on an ongoing basis as part of the range of supported Grid applications. This could be by ARCS, or the MARCS providing the specific compute resources to the ARCS Grid.

We should state that during the 2-year NeAT program we see the above proposed activities as very much complementing the more generic data provision services that AusCover will deliver to users. It is very likely that these more specialised services to the more specialised users might be captured within the body of in TERN 2 (or the follow-on initiative to TERN).

**eResearch effect** *What changes in behaviour and activity are expected from the project that will demonstrate the broader adoption of eResearch practice?*

If the proposed project is effective in delivering the planned services to the targeted users, we anticipate an enthusiastic level of uptake of the products and services developed in this NeAT project.

In the exemplar, we are aware of the interest in and motivation of user communities to directly engage with the utilisation of a new set of eResearch tools to handle large spatio-temporal datasets. We see the immediate need to support this interest with:

- Tools and information, services such as database management software, search services using metadata, processing suites (workflows) that will largely automate the extraction of statistical or other types of user-required information from extensive time series of spatial-temporal products derived from satellite imagery.
- Services that will integrate these with databases of the more traditional agency information to produce increased efficacy of access, enhanced information analyses and beneficial management decision information.

The single greatest impact on the earth system science users will be to combine a processing system (the Grid) with the output of a sensor (the data) to produce an instrument capable of studying the Earth System and its processes in ways not previously possible. It will accelerate research in the natural sciences by enabling hypotheses that previously required enormous resources and weeks or months of processing time to be tested and revised in a fraction of the time. By making such processing easy to use it will act as a magnet for research and development activities in this domain, simultaneously growing the user communities for both the Grid and the data sets. Lastly, by demonstrating workflows for gridded spatial data, it will likely stimulate similar developments for other data types; for example, demographic or epidemiological data, which will lead to data integration, and eventually, domain integration.

**Broader adoption** *Which additional communities, resource providers or organisations would also be expected to benefit from the provision of the same or similar services should the project succeed?*

The proponents of this project have had a history of significant and long term interaction with the earth system science research community, including the ARC Network in Earth System Science (ARC NESS). This is a group of research scientists whose interests span the spectrum from observational data (particularly satellite data) and also numerical modelling. Workshops and seminars with a focus on utilisation of remotely sensed products that we have held to date indicate that there are a wide range of activities into which the Workflows and Grid services developed in this project could expand. For example, soil moisture observations applied to agricultural and catchment research groups, cloud microphysics assimilation into climate models (eg ACCESS), land surface fluxes for timber plantation research, enhanced fire detection etc.

Spatial data (including remotely sensed fields), and spatial model outputs, are widely used within the terrestrial ecosystem research community, as well as more broadly. The interface/interoperability standards (formats, protocols, metadata services etc) adopted and developed in the course of this project will make it possible for others to add further services beyond the basic suite proposed here, thereby enriching the pool of operations available to the community at large. By providing a route to process (utilise) such data sets, it will also encourage provision of similar data via the TERN/AusCover mechanism.

Expanding the uptake of the range of services developed in the project will require a significant outreach effort. We see this developing more in year 2 of the project when a meaningful level of reportable progress has been achieved. Mechanisms for progressing the wider uptake would be:

- Promotion of the aims, objectives, strategy, progress and planned outcomes at conferences, workshops, refereed literature etc.
- Visits, either by invitation or by deliberate targeting, to organisations where we believe the NeAT program approach would be of interest and direct benefit. There would be a requirement in promoting such activities to demonstrate aspects of the NeAT project achievements, to identify benefits to the participating research groups and agencies (both with respect to services and perhaps economic benefits, if they can be quantified), the enhanced quality of the information produced and user agency endorsements.

The workflow system deployed for this project will be a generic service that could be used by researchers from any discipline who want to run workflows utilising the computing and data services supported by ARCS. This project will therefore provide an excellent exemplar application for a much more general workflow service.

### **Value adding**

*Identify the components of the project that could be based around generic technologies or be implemented through shared services for which the project would provide an exemplar use case or requirement set.*

The raw data products utilised in this proposal will be sourced primarily from the TERN/AusCover system (funded for 3 years, viz. July 2009 - June 2012), including the WASTAC Satellite Data Archive (WASTAC manages the reception and archiving of environmental satellite data downlinked in WA - [www.wastac.wa.gov.au](http://www.wastac.wa.gov.au) ). One of the project proponents, Prof M Lynch, is currently Chair of WASTAC.

The distributed data architecture pioneered in Australia by the AO-DAAC and to be expanded upon by AusCover will underpin the deployment of distributed remote sensing processing on the Grid.

Some of the operations that are envisaged are already implemented and are either embedded in existing processing systems or available as standalone applications. Rather than re-implementing them, these would be abstracted and encapsulated for use in workflows that would run on the Grid.

There is a great deal of work already underway globally to develop workflow description languages and other tools to support workflow management and operation. This project would begin by seeking to capitalise on this work from the outset, to build upon it, and to contribute back to the community.

### **Standardisation**

*Describe the global technology development or standardisation work that will be adopted, adapted or extended within the project and any risk reduction available by collaboration with similar activities occurring elsewhere in the world.*

Existing well-established open data formats and protocols that are well adapted to the heavy-lifting anticipated in this project, such as netCDF, HDF and OPeNDAP, are a natural fit. They are also more flexible than the OGC protocols (WMS and WCS) and better suited to working with real data (rather than GIS objects), though there may be opportunities to leverage other OGC components such as the Catalog Service (eg. as proposed in the NeAT Spatial Information Services Stack project). Spatial metadata standards such as ISO 19115 are also obvious choices, especially as they are already widely mandated. The SISS NeAT project, IMOS eMII and the MACDDAP NeAT project are working on developing tools to support standard spatial metadata and data formats, and close interaction with these projects on this area will occur.

The project will utilise the ARCS National Grid, which is based on international standards from the Open Grid Forum and standard grid middleware, including the Globus Toolkit and VDT, web services, and data transfer services.

There are many existing workflow tools that are available. Commonly used approaches include Kepler and Taverna. ARCS will review existing tools and make a choice of a workflow system that it is willing to support for use in this project and more broadly, based on input from NeAT and other eResearch experts and workflow users, and the requirements of this project.

The data delivery aspect of this project will be aligned with the “standard” data formats of the end users. These data formats are typically GIS compatible, but specific details will be determined through feedback from end users. Data delivery systems will translate from data archive standards to various end user formats (most likely as a standardised workflow component).

## **Project Scoping**

**Key Participants** *Name any PfC components, any NCRIS capabilities, or any other institutions or groups that will need to be involved in the project planning and execution.*

CSIRO Marine and Atmospheric Research, the Bureau of Meteorology, Geoscience Australia, Curtin University, and through them, iVEC, are key AusCover partners that would need to be involved.

ARCS and ANDS, as operators of the Grid and developers of the data commons respectively, are also essential participants. While external groups such as state and federal agencies and universities are not formal participants, they will be key to development and planning in terms of providing the end-user feedback to the project.

The key resource providers for the project are:

- TERN/AusCover, which will provide the remotely sensed products.
- Australian Research Collaboration Service (ARCS) which will provide advice on IT services and also will provide the computing infrastructure on which the project primarily will be undertaken.
- WASTAC, the provider of a locally down-linked satellite data archive with 29 years of observations.

A Project Committee, which provides project governance and steering will be established.

The project will also have a broader Advisory Committee to provide expert advice on the planning and execution of the project.

A project manager will be appointed and will work on the project at least half-time, and will be NeAT funded.

**Proposed Project Committee** (some nominees have not yet been approached)

Professor Andy Pitman, University of NSW (Chair, ARC NESS)

Dr Edward King, CSIRO Marine and Atmospheric Research

Professor Mervyn Lynch, Curtin University

Professor Tom Lyons, Murdoch University

ARCS Executive Director, Tony Williams, or delegate, currently Paul Coddington

ANDS Executive Director, Ross Wilkinson, or delegate, currently Andrew Treloar

Project Manager (ex officio) - to be appointed

## **Project Scale**

*Identify the overall scale expected in the project, eg. 1 to 3 years, total effort in any year, and nominate any parties that have indicated a willingness to participate through providing resources. (funded or in-kind, people or facilities).*

Conduct of the project described will require funding for 2-3 FTEs per year for the two years.

In-kind resources contributed (not requiring NeAT funds):

- Considerable data storage infrastructure and data will be brought to the project by AusCover participants, together with expertise in the application software that is to be made available via the workflows. WASTAC (satellite data archive), NCRIS TERN (AusCover products), grid services and data services (OpenDAP servers) and support staff from ARCS, computing and data storage infrastructure and services from MARCS such as iVEC, TERN data.
- Curtin University, CSIRO and Geoscience Australia (office space, staff supervision, project management, coordination of user workshops, project reporting).
- AusCover is likely to devote between 1 and 2 EFT to architectural tasks, at least for the first two years, and it is most likely that some fraction of those would be contributing directly to this project.

## **Major Steps**

*Identify the key steps that will be visible to users as the services develop. Note that some observable deliverable is needed every half year and projects may be reviewed based on the achievement of these steps.*

**6 months** – Hold a research users workshop and work with the Project Committee and AusCover to refine the NeAT project and define user needs and research priorities, and decide on initial data processing components to be supported, and on some exemplar workflows for implementation. Interact with ARCS and ANDS to establish the Workflows strategy. This workshop will require some initial work on demonstrating trial workflow processes to have been completed in order to test feasibility. These should include integration of existing processing applications and processing chains into a Grid-based data processing and delivery system. A decision on a workflow system to use, and trialling it with simple test workflows utilising ARCS grid computing and data services could then follow

**12 months** – User Workshop #2. Advice to and feedback from the interested research groups. Demonstrated data access using standard ARCS services and initial availability of prototype Workflows implementation. Techniques for abstracting and encapsulating services and passing intermediate results between them within the workflow system will be demonstrated. Some workflows developed for AusCover data providers to generate standard derived data products in standard data formats from selected raw data sets. Design of some exemplar end user workflows as processing chains of disaggregated components, *ie.* Decompose existing processing chains, where appropriate, to develop workflow and interfacing methodologies.

**18 months** – User Workshop #3. A further round of interactions with the research user groups. With end user feedback, and through working collaboratively with a small number of end users, we will further enhance the data access and Workflows tools to include data manipulation, integration of agency data and integration of statistical analysis tools. Development of prototypes of exemplar user workflows.

**24 months** – User Workshop #4 and final round of interaction with research groups. End users will be able to access, manipulate and analyse AusCover (plus other) remotely sensed data and products using a functional Workflows-based systems. Some exemplar workflows developed for some end user groups.

**Dependencies**

*Identify dependencies that exist to activities or developments external to the project.*

This project depends on data being made available by the TERN/AusCover project, and/or the WASTAC and Curtin University data archives at iVEC. It is expected that there will be dependencies on existing infrastructure at iVEC, ARCS Grid services, and technologies developed within the AO-DAAC project. We expect minimum project risk to arise from these collaborations because both Curtin and CSIRO are participants in TERN/AusCover, both are members of the WASTAC consortium and both are members of iVEC, which is a member of ARCS. CSIRO is already collaborating with ARCS to make sample data sets available online.

\*\*\*\*\*