19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

TERN/AusCover - Remote Sensing Data Management for Terrestrial Ecosystem Research

M. Paget [a], E. King [b], L. Edwards [c], M. Wyatt [c], R. Jones [d], M. Gray [e], K. Johansen [f], I. Grant [g] and A. Held [a]

[a] *CSIRO Marine and Atmospheric Research, Canberra, ACT.*
[b] *CSIRO Marine and Atmospheric Research, Hobart, Tas.*
[c] *iVEC, Kensington, WA.*
[d] *CSIRO ICT Centre, ANU, Canberra, ACT.*
[e] *Department of Imaging and Applied Physics, Curtin University of Technology, Perth, WA.*
[f] *Centre for Spatial Environmental Research, University of Queensland, Brisbane, Qld.*
[g] *Data Quality and Improvement Section, Bureau of Meteorology, Melbourne, Vic.*
Email: matt.paget@csiro.au

Abstract: AusCover (www.auscover.org.au) is a facility of the Terrestrial Ecosystem Research Network providing a national expert network and data delivery service for Australian biophysical remote sensing data time-series, continental-scale map products, and selected high-resolution data sets. There are ten partner agencies, located in all the Australian mainland capitals, participating in the AusCover program. The AusCover data system underpins the facility by providing enabling infrastructure for geographically distributed management and sharing of large data sets and by presenting a unifying view of the data collections through a virtual data centre. The data system addresses three of the main problems encountered by non-expert prospective users of remote sensing data: what data exist, where are they, and what do they mean? Simultaneously, it seeks to assist data producers and providers by providing frameworks, protocols and tools that make it easier to organise, describe and serve data sets. The volume of modern remote sensing data sets combined with that of historical archives, together with the range of custodians and providers, means that a distributed approach to data storage is essential. In order that diverse data sets can be identified, located and accessed within such a system, the adoption of open standards for describing and representing data has been critical to the development. For high-level descriptions of data sets, the ANZLIC profile of ISO 19115 was chosen so as to 1) simplify the metadata, 2) enable the use of existing tools such as Geonetworks and ANZmetlite for management, and 3) maximise the chances of its discovery through related data catalogues. An effort has been made, wherever practical, to store data sets in self-describing and architecture independent file formats such as NetCDF and HDF. This confers the advantages of easily storing data and metadata (and provenance information) together, allowing uniform access across platforms, and providing for a future migration path that can be fully automated. AusCover does not however require the use of these formats; if it is impractical to store data this way, AusCover can still host the data files and a metadata record to enable data discovery, or even only the metadata record if custodians require that the data be stored outside the AusCover system. The use of open file representation standards permits several value-adding services to be built on top of the data system. Advanced data server software, such as THREDDS and the OPeNDAP Hyrax systems, support the OGC web service protocols and the Data Access Protocol (DAP) on NetCDF, HDF and some other format files. The exposure of file-level metadata by these servers enables automatic harvesting of granule metadata for catalogue population and indexing. Additionally, the OGC and DAP protocols support user-customised on-demand subsetting and representation of data, both for visualisation and delivery. On top of the physical data system and server infrastructure, the AusCover Data Systems team is developing frameworks and protocols to assist management of the data content. The team provides guidance on data formatting and organisation, licencing and custodianship responsibilities. By working with the AusCover Data Products team, guidelines are being developed for documentation and quality assurance and are being incorporated into a governance framework, which defines the processes and mechanisms by which data are managed and the roles and relationships of the various participants. The governance framework reduces the data discovery, access and interpretation burden on users by ensuring consistency across the data collection.

Keywords: *Satellite Remote Sensing, Data management, Web services, System integration*

1. INTRODUCTION

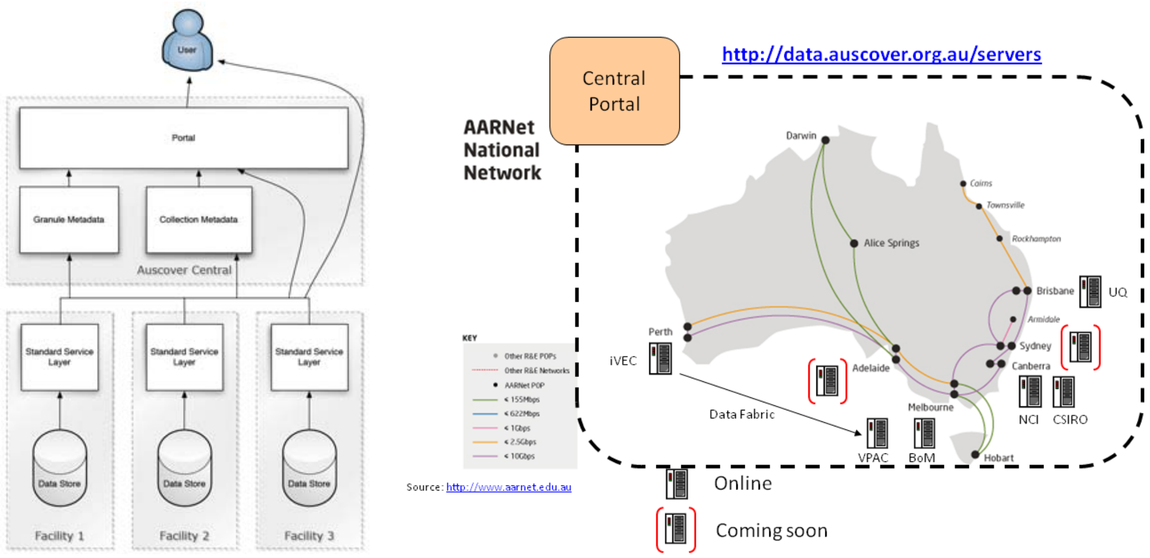
The Terrestrial Ecosystem Research Network (TERN) is an NCRIS/EIF funded program designed to bring Australian ecosystem researchers together and to foster collaboration through collating, calibrating, validating and standardising existing ecosystem data sets and research activities. TERN facilities focus on ecosystems at a range of spatial scales from intensively monitored local sites through regional transects to generation of continental-scale maps of biophysical properties. The AusCover facility covers the latter. Key sources of continental-scale data are satellite remote sensing data, synthesis of regional data into national data sets and environmental process modelling. AusCover provides national coordination and improves collaboration between local, regional, state and national remote sensing facilities around Australia. The AusCover partner agencies are listed in Table 1. These partners represent the experts in remote sensing data management, processing and analysis in Australia. CSIRO Marine and Atmospheric Research, Canberra is the lead agency for AusCover.

Table 1. AusCover partner agencies with a brief comment on their area of expertise relevant to AusCover.

Agency	Specialty
CSIRO Marine and Atmospheric Research, Canberra	Continental time series, data management and operational system development
Geoscience Australia	Data archives, ground stations and international agency coordination
Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES) and Dept. of Climate Change and Energy Efficiency	National integration of environment surveys and monitoring activities
Joint Remote Sensing Research Program (University of Qld, Qld Dept. Environment and Resource Management, NSW Dept. Environment and Heritage)	Field validation of remote sensing data
Bureau of Meteorology	Operational delivery of products linked with weather and climate
Curtin University and iVEC	Satellite remote sensing data processing, and metadata standards and systems
Charles Darwin University	Fire and burnt area monitoring
University of Adelaide	Rangeland and desert monitoring
University of Technology Sydney	Plant phenology metrics

The AusCover data system underpins the facility by providing a distributed, yet unified, set of data servers and web-enabled data delivery services. The distributed servers, along with an early decision to promote the use of open-source or freely-available software, ensures that the system as a whole is flexible with regard to the location and number of servers and the web services that expose the data. Presently, data sets are hosted at large data centres including CSIRO Data Centre and ANU NCI in Canberra, ARCS Data Fabric, Bureau of Meteorology and University of Queensland (Figure 1). More servers are scheduled to come online in the coming year. Each data server has a selection of web-enabled data services attached to it to allow open archive browsing and free file download. The primary data sets that AusCover provides access to are raster-based gridded remote sensing products. These data sets typically contain many hundreds or thousands of files, each of which can be megabytes to gigabytes in size. Taking a lead from the oceanographic and climate and atmosphere communities, the primary file formats that best suit large volume gridded data are NetCDF, HDF and GeoTIFF and the primary web-enabled data server software packages are OPeNDAP, THREDDS and Geoserver. The AusCover system aspires to present the majority of its data in at least one of these formats and via at least one of these services. However, AusCover is not prescriptive in this. The minimum requirement to join a data set to the AusCover system is to create a metadata record so that the data set can be discovered.

The AusCover data system is not just a set of distributed servers. Centralised services include advice and support for creating metadata records, tools for file format translation, a data base of metadata records, detailed product information and a discovery and visualisation interface. Central coordination also involves active engagement with national and international activities of a similar nature; that is, metadata and data standards for web-enabled data delivery.



Source: http://www.aarnet.edu.au

(a) (b)

Figure 1. (a) The distributed servers form a virtual network unified by common central data access and discovery tools based on standard protocols. (b) Servers are physically connected to the AARNet network.

2. PRINCIPLES OF DESIGN

The AusCover system is based on a set of data management principles that recognise practical considerations, as well as the long term nature of the NCRIS program and the wider needs of TERN itself.

2.1. Data Sharing

The value of data lies in its interpretation by the user. By making data easy to share and use its value can grow many times over as it is repurposed repeatedly. For large data sets with complex processing, the investment in production is substantial so the gains that can be realised by sharing and re-use are often great. Moreover, in a paradigm of environmental change, data can often increase in value as it gets older, so preparing data so that it can be shared with researchers in the future is also profitable. The additional short-term work that is required to preserve data is therefore likely to be paid back over the longer term.

2.2. Standards and Interoperability

Consistency in representation and description of data sets is important because it enables tool reuse and, hence, flexibility in processing methodologies. It also allows human users to quickly understand new data sets that are stored in familiar ways. Consistency in organisation and labelling opens the door to automation so that the need for manual handling in the future is reduced. Consistency also supports the use of standard interfaces that enable data to be accessed easily across different processing platforms and toolsets, permitting a much wider range of reuse and sharing and leading to innovative use in new domains and synthesis studies.

2.3. Metadata

Data never stands alone. Metadata describing how data was obtained, stored, processed and organised are essential to enable its interpretation without ambiguity. For AusCover, where data sets often comprise many similar files, we distinguish between two different types of metadata: Collection and Granule. Collection attributes apply across the whole of the data set, such as sensor and processing method, general space and time extents, creator and custodian, references and quality assurance information. Granule attributes are specific to and vary with each image (granule), such as specific location or space range and time of acquisition. Collection metadata describe properties of the data set that are generally static and are often useful to discover data sets. In contrast, the granule metadata vary across the collection (*e.g.* spatial coverage and acquisition time for a scene) and are key for locating and making use of individual scenes (in conjunction with the collection metadata).

2.4. Formats

Format standards that implement a well defined data model, support metadata storage, and provide documentation and tools for accessing them across a range of platforms facilitate data use by supporting correct access and interpretation over time and are independent of the usage environment or context. Having data access and its interpretation tied to individual researchers and their specialised lab software is a recipe for exclusivity and a loss to the wider research community. Data models are an abstraction that permits a consistent conceptual representation of both data and metadata to users that can be separated from the storage organisation. Toolsets exist that provide users with access to data and metadata objects stored in files that are well documented, open source, and that work across multiple platforms and languages. AusCover encourages and supports the use of these types of formats but it takes the view that if a data set is organised in a consistent format that is sufficiently open as to be future-proof then it can be included.

2.5. Machine Readability

Providing data in standard formats with standardised metadata means that machine readability is possible, leading in turn to automatic use, including future format migration. Automating access is not only a great labour saving, permitting data to be reused efficiently and frequently, but it also enables machine migration of data to new formats or storage mechanisms in the future.

2.6. Licensing

Data licences can be important to protect the IP of data producers and to control the uses data are put to. While AusCover seeks to be guided by the NCRIS principles of open access to data for research purposes, it is recognised that some data sets are necessarily encumbered by licences that would prevent them from being available at all were they not respected. TERN is in the process of defining a data licence policy for use across all facilities, and AusCover will ultimately be guided by this policy. The present default choice is Creative Commons Attribution (CC-BY) for data sets hosted by AusCover (http://creativecommons.org/). Where a restrictive licence is required, AusCover recommends that the data provider self-host and manage user access to the data set, and AusCover will provide a metadata record to support discovery. AusCover expects to support and encourage the use of data citations in the future.

2.7. Provenance

Provenance is essential metadata that supports interpretation by making it possible to understand how the data was obtained and modified and, therefore, how it represents what was originally measured. The simplest form of provenance metadata is to record version numbers for both input data sets and process steps, allowing manual traceability. More advanced provenance systems codify artefacts, processes and agents, and the relationships between them, enabling automation of traceability. Provenance is very important in remote sensing where sensors themselves are complex, and the data are then subjected to many different processes that may depend on parameters that evolve and are revised over time. Provenance information for each AusCover data set will be collated in a product user webpage while we identify the important elements that may be standardised or codified.

2.8. Governance

A governance framework is essential for data management because it provides a consistent way of understanding the state or condition of individual data sets in a broader context. For example, in AusCover, a data set may pass through a series of quality assurance and reformatting steps so that once it reaches a particular state the end user can be sure that particular metadata and storage conventions will be in place. In this way a governance framework provides consistency and helps to build user trust through familiarity. AusCover will rely on a clear and adaptable framework for success as it builds its network of data providers and data management experts.

3. IMPLEMENTATION

The AusCover data system was made “live” in March 2011. The live release demonstrates searchable collection metadata records, open data formats with embedded collection and granule metadata, and programmable or manual web-enabled data access.

3.1. Metadata

AusCover has adopted the ANZLIC profile for collection metadata (http://www.osdm.gov.au). It provides ISO compliance, it is very general (inclusive) and it is in widespread use in the terrestrial, marine and climate domains in Australia and New Zealand. For granule metadata, the Climate and Forecast (CF) conventions (http://cf-pcmdi.llnl.gov/) provide metadata standards that are particularly suited to gridded environmental data. CF originated from the oceanographic and atmosphere communities where data interoperability was important for data assimilation and earth-system and climate modelling activities. CF provides a basis for good data management particularly with respect to labelling and arranging of dimensions and variables and a nomenclature to describe how data in a grid cell or at a point were generated. Not all AusCover data files are expected to be CF-compliant. The goal is to provide expertise and tools that can enable data providers to migrate their data files towards CF-compliance, or other suitable standards as they emerge.

3.2. Data formats

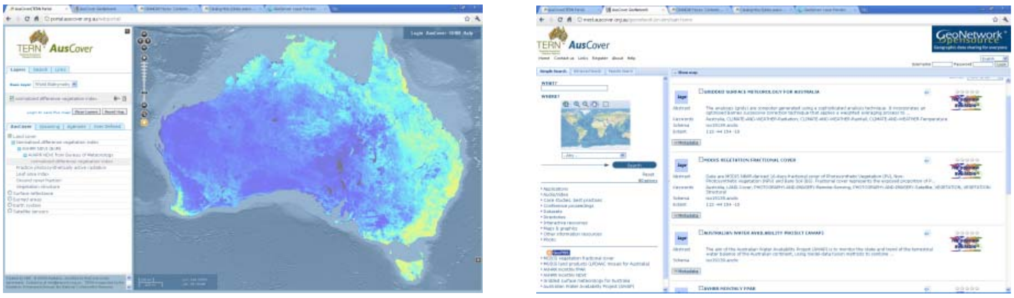
Historically, many different data file formats have been used to store remote sensing and related field measurement data. Proprietary file formats, where a specific piece of software or a paid licence is required to read the data, are a barrier to interoperability. In contrast, data in open (non-proprietary) file formats can be accessed by a variety of tools and the choice of tool is with the user rather than with the data provider. The file formats recommended by AusCover for gridded data are currently NetCDF (http://www.unidata.ucar.edu/software/netcdf), HDF (http://www.hdfgroup.org), GeoTIFF (http://trac.osgeo.org/geotiff), and spatial database data or (ESRI) shapefiles (http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf), which can be ingested into a spatial database. These file formats are in common use within the remote sensing, geographical and geospatial communities; are recognised by the Open Geospatial Consortium (OGC), and; are supported by web-enabled data servers that provide value-added services such as sub-setting, aggregation and visualisation that improve the usability of data. Other file formats can be served from AusCover systems but likely only as whole files (the traditional access method, eg FTP). A file format of NetCDF, HDF, or GeoTIFF with embedded CF metadata allows the greatest range of access methods because the format and metadata are standards-compliant and many tools are available, or can easily be developed, to read these files in server-side or desktop environments.

3.3. Data services

The advent of portable, self-describing, structured data formats has in turn enabled the development of advanced data services which are able to take advantage of the additional information provided about data sets to return specific subsets rather than whole files. The main protocols supporting this type of access are the Data Access Protocol (http://www.opendap.org) and the OGC Web Protocols (WFS, WMS and WCS, http://www.opengeospatial.org/). The OPeNDAP Hyrax, THREDDS (TDS) and Geoserver data servers variously support these protocols for suitably formatted and described data sets. A particularly powerful feature of Hyrax and TDS is their ability to use server-side catalogues that provide extra metadata to the served files so that, when accessed via the enhanced data protocols, the files appear to the user to contain the additional metadata. This provides a very efficient means of adding collection metadata to a large data set. In addition, for data files that do not conform to the format and metadata conventions, both servers provide whole-of-file access via plain HTTP, in a similar manner to an Apache HTTPD, thereby catering to the needs of traditional file-at-a-time data users. AusCover data sets will be exposed by these services (that are relevant for each dataset) to allow both visualisation (e.g. OGC) and programmatic access (e.g. subsets).

3.4. Central Services

The distributed data servers, all using common protocols and hosting data sets that fully or partially conform to a set of metadata protocols, are unified by a group of tools that provide cataloguing and a single point of data discovery and access. The two key services are the visualisation portal and Geonetworks (Figure 2). The portal, which is currently based on the Integrated Marine Observing System (IMOS) Portal, provides a simple menu-based and map-driven tool for exploring data sets within the data system. Geonetworks is an open source ISO19115-compliant metadata repository and search tool. The Geonetworks system provides a master catalogue of AusCover collection metadata (i.e. metadata describing data sets as a whole). It is the starting point for all query–based searches for AusCover data. Once data sets have been identified, the prospective user can navigate directly to the web interface of the relevant data server or examine them via the graphical portal interface. A future development will permit fine grained searching for specific granules within a data set based on temporal and spatial criteria. The AusCover data systems website (http://data.auscover.org.au) contains links to the Geonetworks, the portal, the servers and the data products. An automated tool has been implemented to check the current online status of each of the servers within the AusCover data system. A list of currently available and planned data products is maintained in conjunction with the AusCover data products team. In development are a set of product user pages that will give extended information about each data product beyond that traditionally available in a structured metadata record, such as an easily-understood product description, algorithm details, scientific applicability and limitations, examples of use, provenance and other "free text" information.



(a) (b)

Figure 2. (a) The visualisation portal (b) and Geonetworks metadata server are the two most visible tools.

4. GOVERNANCE AND WORK FLOW

Having established a network of physical servers and a software framework of protocols, formats and conventions, the AusCover data system is at the point where it is ready to accept data files and metadata records. Ensuring consistency and completeness in a distributed system, especially one maintained by multiple individuals or groups, is a challenge. In order to capitalise on the advantages conferred by the adopted standards and the consistency they confer, it is essential to have in place some conventions that ensure that they are adhered to. For example it is important that every data set has, at the very least, a collection metadata record in the central Geonetworks catalogue and a suitable descriptive page on the AusCover web site providing the background information about the data set. Each of these elements are required to be in place in order for the three basic data questions (what exists, where is it, and what does it mean?) to be answered. The set of rules, roles and processes put in place to manage this problem constitute a governance framework.

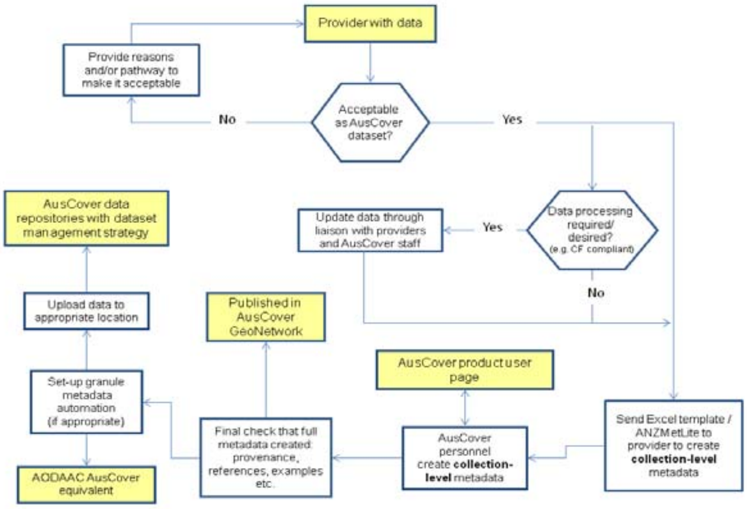

Provider with data
Provide reasons and/or pathway to make it acceptable
No
Acceptable as AusCover dataset?
Yes
AusCover data repositories with dataset management strategy
Update data through liaison with providers and AusCover staff
Yes
Data processing required/ desired? (e.g. CF compliant)
No
Upload data to appropriate location
Published in AusCover GeoNetwork
AusCover product user page
Set-up granule metadata automation (if appropriate)
AODAAC AusCover equivalent
Final check that full metadata created: provenance, references, examples etc.
AusCover personnel create collection-level metadata
Send Excel template / ANZMetLite to provider to create collection-level metadata

Figure 3. Flowchart illustrating the pathways and processes to be followed when incorporating new data into AusCover. The workflow begins with an iterative process between the data provider and AusCover to ensure the dataset meets the AusCover goals. Data processing (e.g. to CF compliance) is an optional step that can be returned to after a metadata record has been created and published. An Excel template or the ANZMetLite software is used to collate the required metadata, after which AusCover personnel create an ANZLIC-compliant and AusCover-consistent metadata record.

The governance model for AusCover is still emerging as we begin populating the data system. It identifies roles for data providers, data custodians, metadata system administrators and the data products team. The set of processes are represented as a workflow in the flowchart in Figure 3. The issue of data set versioning, critical to correct identification, also falls under the umbrella of system governance, as does the provision of a licencing framework. Whilst the governance is largely implemented in the context of the data system, the AusCover data products team plays an important complementary role. They have responsibility for ensuring that the data sets submitted come with adequate metadata and supporting descriptive material to enable correct discovery and interpretation of the data.

5. CONCLUSION AND NEXT STEPS

The basic AusCover data system, comprising the key central portal tools and several data servers is online. The immediate task is to grow the data set collection hosted by the system and, in the process, refine the governance model. Simultaneously, to support the increased number of data sets, more data servers will be established at each of the participating nodes. Testing is also underway to ensure that discovery and harvesting of AusCover collection metadata records by the central TERN portal is possible. This feature is essential to ensure that AusCover integrates into TERN at large, and beyond that, with the other related environmentally-focussed NCRIS capabilities, including particularly IMOS, the Atlas of Living Australia and Research Data Australia.

There are a number of areas where further development is planned:

- Automated granule cataloguing – the data sets comprise (often) large numbers of images. It is intended to utilise the AODAAC tool (http://espace.library.uq.edu.au/view/UQ:155380) developed within IMOS to crawl the data server URLs to identify all data granules and harvest metadata from within them. This will permit the generation of customised queries of the data servers that return only those subsets of the data that users require.
- Tools to assist data providers – it is recognised that not all data providers will have the capacity to reformat and document their data to the AusCover preferred standards. While these providers are catered for by the lowest rungs of the AusCover data acceptance policy, we intend to develop portable toolsets to assist in the reformatting and preparation of data sets for inclusion at the highest standards of conformance. This will also include development of documentation templates to guide and ease the contribution of well-described data to AusCover.
- Provenance information – remote sensing products usually result from a multi-step processing pipeline that contains many uniquely configurable steps. The choices made in acquiring and processing data constitute information about its provenance that frequently have a direct bearing on its interpretation and hence utility for particular scientific usage. Automated management of provenance information is an area of data informatics that is rapidly developing and we will monitor it for tools and techniques that can be exploited by AusCover.

At a minimum the development of AusCover will result in many existing and new data sets being stored in portable self-describing file formats with standardised metadata. This single feature is a giant step towards protecting these valuable national data assets for use and reuse. If, at some point in the future, the choices made in this aspect of the AusCover system prove to be inadequate for future needs, the task of reformatting the data will be able to be automated. Beyond this the adoption of widely used open standards will open the gateway to integration of remote sensing data into the wider fabric of spatial-temporal environmental data in Australia. The ultimate indicator of success for the AusCover system will be that it makes data so easy to contribute, discover and store, that it becomes the preferred model for archive within the research user community.

ACKNOWLEDGEMENTS

AusCover is a facility of the Terrestrial Ecosystem Research Network, which is supported by the National Collaborative Research Infrastructure Strategy and the Super Science Initiative.

REFERENCES

Acronyms and technologies are referenced in the text with a corresponding URL web address.