GeoResJ 6 (2015) 135–140

Contents lists available at ScienceDirect

GeoResJ

journal homepage: www.elsevier.com/locate/GRJ

Unlocking the Australian Landsat Archive – From dark data to High Performance Data infrastructures

Matthew B.J. Purss [a,*], Adam Lewis [a], Simon Oliver [a], Alex Ip [a], Joshua Sixsmith [a], Ben Evans [b], Roger Edberg [b], Glenn Frankish [c], Lachlan Hurst [d], Tai Chan [e]

[a] Geoscience Australia, Australia
[b] National Computational Infrastructure, Australia
[c] Lockheed Martin Australia, Australia
[d] Victorian Partnership for Advanced Computing, Australia
[e] Cooperative Research Centre for Spatial Information, Australia

ARTICLE INFO

Article history:
Received 10 October 2014
Revised 17 February 2015
Accepted 19 February 2015
Available online 27 March 2015

Keywords:
Big data
Landsat
High Performance Data
High Performance Computing
Data rescue
Earth Observation

ABSTRACT

Earth Observation data acquired by the Landsat missions are of immense value to the global community and constitute the world's longest continuous civilian Earth Observation program. However, because of the costs of data storage infrastructure these data have traditionally been stored in raw form on tape storage infrastructures which introduces a data retrieval and processing overhead that limits the efficiency of use of this data. As a consequence these data have become 'dark data' with only limited use in a piecemeal and labor intensive manner. The Unlocking the Landsat Archive project was set up in 2011 to address this issue and to help realize the true value and potential of these data.

The key outcome of the project was the migration of the raw Landsat data that was housed in tape archives at Geoscience Australia to High Performance Data facilities hosted by the National Computational Infrastructure (a super computer facility located at the Australian National University). Once this migration was completed the data were calibrated to produce a living and accessible archive of sensor and scene independent data products derived from Landsat-5 and Landsat-7 data for the period 1998–2012. The calibrated data were organized into High Performance Data structures, underpinned by ISO/OGC standards and web services, which have opened up a vast range of opportunities to efficiently apply these data to applications across multiple scientific domains.

1. Introduction

The Landsat mission is the longest continuous civilian environmental observation and monitoring program in history [1]. It has been continuously acquiring Earth Observation data dating from the launch of Landsat-1 in July, 1972 to the present day; a period of 42 years. In many cases the Landsat archive has provided the only consistent source of information to monitor changes in the surface of the Earth [2]. Following a change in data policy by the US Government in 2008 the Landsat archive has been made open and freely available to the public over the internet [3,4]. With the successful launch of the Landsat-8 satellite in February 2013 this mission is set to continue providing valuable contribution to society for many years to come.

While the sheer volume and richness of the data acquired by the Landsat mission provides immense value to the scientific community, it has historically presented a major challenge and barrier to the effective use of this data to contribute to global environmental and ecological research initiatives. Traditional approaches to the use of this data have focused on processing and analysing individual scenes of data as a series of snapshots of the surface of the Earth. This has proven to be a very inefficient and labour intensive way to use this data to inform Earth Observation analyses. Also, because of the costs of data storage infrastructure, this data has traditionally been stored in raw form on tape storage infrastructures which introduces a data retrieval overhead that limits the efficiency of use of this data.

The result has been that the Landsat archive has been 'locked' up within these tape storage infrastructures hosted by various Government Agencies and often not used to contribute to environmental and Earth Observation research initiatives because of the difficulties in accessing and working with the data. This has significantly reduced the impact and application of this data and resulted in this data becoming 'dark data'.

* Corresponding author.

http://dx.doi.org/10.1016/j.grj.2015.02.010
2214-2428/

To address this issue, the Australian Space Research Program funded the Unlocking the Landsat Archive (ULA) project; a 3½ year, AUD$3.5M public/private consortium project that operated from April 2011 to July 2013 [5,6]. The fundamental aim of the ULA Project was to improve access to Australia's archive of Landsat data, and provide an analysis capability for delivery of environmental information to inform and support government policy [7]. The ULA Project was led by Lockheed Martin Australia (LMA) and involved technical input from Geoscience Australia (GA), the Victorian Partnership for Advanced Computing (VPAC), the National Computational Infrastructure (NCI) at the Australian National University (ANU) and the Cooperative Research Centre for Spatial Information (CRC-SI).

2. Rationale

There have been seven functional Landsat satellites spanning from 1972 to the current Landsat 8 mission, constituting the longest running civilian enterprise for acquisition of satellite observations of the Earth. These multispectral images are a unique resource for global research and applications in agriculture, cartography, geology, forestry, regional planning, surveillance, education and national security. Australia has participated in the Landsat program since 1979, when it established the Alice Springs Landsat Downlink Station in Central Australia, and has collected data from nearly every pass of successive Landsat missions over the Australasian region ever since.

As part of the Australian Space Research Program, the objective of the ULA Project was to make the wealth of knowledge available from Australia's Landsat holdings to the community in a well-defined and sustainable manner. The technical work program under the ULA Project was run in four (4) parallel streams, as follows:

1. *Earth Observation Science.* This stream of development was to improve the fundamental processes and algorithms that are used to transform the raw Landsat images into useful Earth Observation products (see Figs. 1 and 2). On a technical level this included geometric correction and ortho-rectification (using the Level 1 Product Generation System [LPGS] provided by the US Geological Survey [8]), spectral calibration to produce sensor and scene independent surface reflection observations [9–11], implementing cloud and cloud shadow detection algorithms [12,13] to produce an improved Pixel Quality Assessment Product [14] and the implementation of a fractional cover classification algorithm [15,16] to produce an information product to enable direct assessment of bare Earth, green and non-green vegetation cover.
2. *National Nested Grid.* The adoption of a unified and common geospatial framework for the interoperability of raster data sets at different resolutions is a significant step in the reduction of the complexity and efficiency of accessing Earth Observation data. This project undertook the development of the National Nested Grid (NNG) concept (see Fig. 3) and worked with the Australian and New Zealand Land Information Council (ANZLIC) to publish the NNG as an ANZLIC Specification Guideline [17].
3. *Workflow Development.* Traditionally satellite data processing has taken place in a very linear workflow with standard techniques for creating the relevant metadata and media distribution [4,18,19]. This project enhanced these capabilities by implementing the NNG concepts as part of the satellite data processing workflow and bridging the gap between the current satellite processing capabilities and the large scale processing that is required to provide full access to the Landsat time series. This was done with a view to extending the framework to other large satellite data collections and includes measures for provenance and quality at the pixel level (see Fig. 4).
4. *High Performance Computing.* Processing large satellite data archives requires significant computing resources. This need has been steadily increasing with each new Earth Observation satellite launch. This project explored and prototyped appropriate computational techniques and scalable resources to underpin Australia's ongoing Earth Observation science capability and capacity. The computing stream implemented the systems developed in the 3 other streams on the NCI.

Fig. 1. Animation of the 25 m Australian Reflectance Grid (ARG25) across the Australian continent showing seasonal variations throughout the year. Continental mosaics were produced using temporal stacks of ARG25 data from 2000 to 2011.

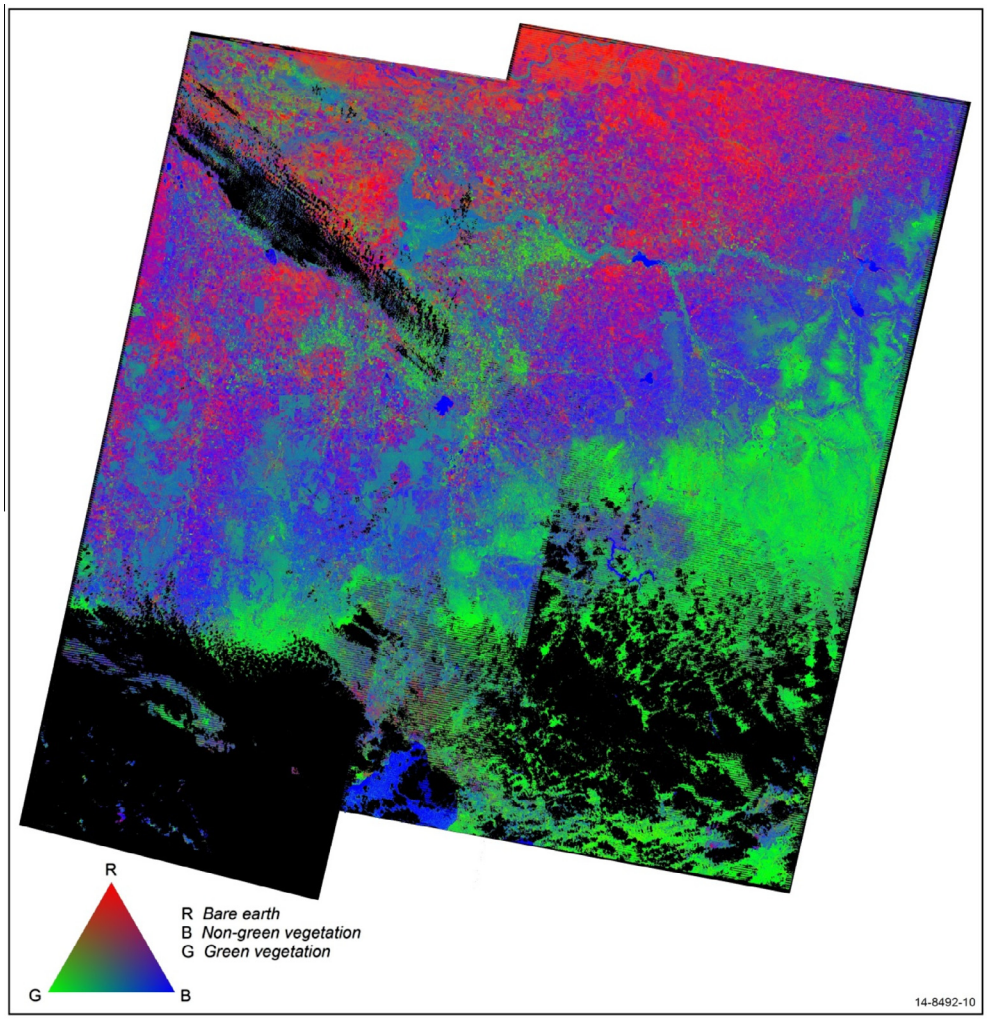

Fig. 2. Four (4) Scene composite RGB image of 25 m Fractional Cover (FC25) from both Landsat-5 and Landsat-7 sensors (Paths 92 & 93, Rows 85 & 86). Red = Bare Earth; Green = Green Vegetation; Blue = Non-Photosynthetic Vegetation. The black regions across the image represent pixels that have been masked out due to the effects of cloud and/or cloud shadow as identified by the Pixel Quality Assessment Product. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Outcomes

The ULA Project has successfully developed Australian capacity in Earth Observation and is an excellent example of how Australia can contribute its niche capabilities for the benefit of Australia and the international community. It has led to several important developments (such as the Australian Geoscience Data Cube – [20,21]) that are enabling government and the private sector organisations to develop effective systems to deliver Earth Observation data for the satellite programs the Commonwealth of Australia participates in internationally. These developments fall into three broad categories, as follows:

(1) The establishment of a spatially aware high volume data processing capability for Earth Observation at the NCI;
(2) The development and utilization of scientifically defensible atmospheric correction algorithms and products for the Landsat satellite sensors in the Australian context (including the systematic identification of cloud, cloud shadow and saturated pixels); and,
(3) The development of scalable spatially-enabled data frameworks and query tools that dramatically reduce the processing time overheads for data queries at continental scale.

In addition to these infrastructure developments the ULA Project has delivered a series of new Earth Observation products derived from Australia's Landsat holdings. The most fundamental of these products is a new surface reflectance product based on industry leading processing algorithms. These data products have been released to the public as the 25 m Australian Reflectance Grid (ARG25) via OGC web services (http://dx.doi.org/10.4225/25/5487CC0D4F40B).

The ARG25 product addresses the atmospheric, sensor specific and terrain effects that make the Level 1 (L1T) Landsat products difficult to work with for large scale spatial and temporal analyses. This has been achieved by implementing a physics based correction to the Landsat data after L1T processing known as the Nadir BRDF (Bidirectional Reflectance Distribution Function) Adjusted Reflectance correction, commonly referred to as NBAR correction [9–11,22]. The calibrated ARG25 product enables users to be confident that the data values for surface reflectance are consistent across the entire data holding and across different sensors having had local variations caused by atmospheric and sensor specific geometry effects removed.

To address cloud effects, and other data quality issues, Geoscience Australia has also developed an advanced "Pixel Quality Assessment" Product [14]. This product creates a pixel level

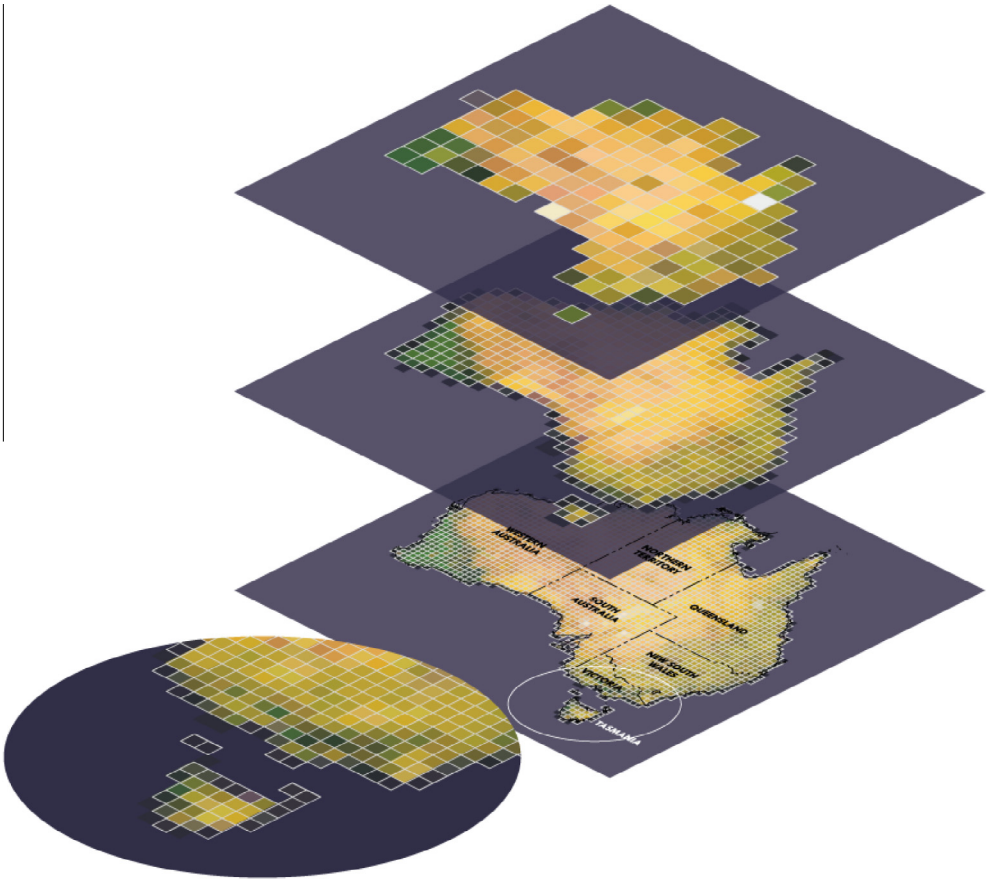

Fig. 3. Conceptual model of the National Nested Grid showing the change in grid mesh sizes with different layers of resolution within the NNG structure.

classification of data quality for the ARG25 product. The scene-independent consistency of the surface reflectance correction allows for the cloud detection algorithms, e.g. [12,13] to be applied in a consistent fashion across all inputs contributing significantly to the accuracy of the product. The generation of an accurate cloud mask enables the remaining "cloud and cloud shadow free" surface reflectance data to be analyzed with confidence.

A fractional cover classification product (FC25) [15,16], has also been implemented using the ARG25 product and the Pixel Quality product as inputs. The FC25 product has similar spatial-temporal coverage to the ARG25 product and is a critical end product for many land management applications; including (for example):

(1) Monitoring crop/pasture condition and the effects of changes in land management practices over time [15,23]; and,
(2) Monitoring for, and mitigating the risk of, wind and water erosion across Australia in response to climate driven events (such as bush fire and flood) [24].

By calibrating and organizing this data into efficient High Performance Data (HPD) infrastructures underpinned by ISO/OGC standards and services the ULA Project has opened up a vast range of opportunities to use these data and combine it with other data sets and data types across multiple domains [20,21,25].

The ULA Project has enabled this archive of data to be used in new ways that allow users to expose and exploit the immense value of the time series of Landsat observations throughout the entire archive at the individual pixel level rather than on an individual scene-by-scene (or image by image) basis [26]. It is now possible to perform statistical analyses of the entire Australian Landsat Archive from 1987 to 2012 for the entire continent (more than 21 $\times$ 10^24 individual pixels) in approximately 6 h. Before the ULA Project this analysis would have taken more than 8½ years to conduct.

The ULA Project has provided a platform that will allow the immense value and impact of the Australian Landsat Archive to be realized. It has indeed 'unlocked' the Landsat archive and is enabling Landsat data to be accessed and used in ways that were not possible just 12 months ago. The archive is now being used to inform research and government policy decisions on a wide range of issues including:

(1) Forestry management;
(2) Surface water and flood impact analyses;
(3) Crop management evaluation and food security assessments;
(4) Forest carbon inventory assessment;
(5) Coastal and shallow water bathymetry studies; and,
(6) Urban planning.

The ULA Project exceeded expectations and provided an invaluable contribution to Australia's environmental research and eResearch infrastructures through the establishment and demonstration of 'data as infrastructure'. The success of the project was recognized in the 2013 International Data rescue award in the Geosciences where the ULA Project received an honorable mention [27,28]. While the ULA Project and its funding has now ceased the Australian Landsat Archive and its associated HPD infrastructures has been accepted into the Research Data Storage Infrastructure (https://www.rdsi.edu.au/) as a dataset of national significance. This will enable this data to be managed and made available to the broader research community into the future in a sustainable fashion.

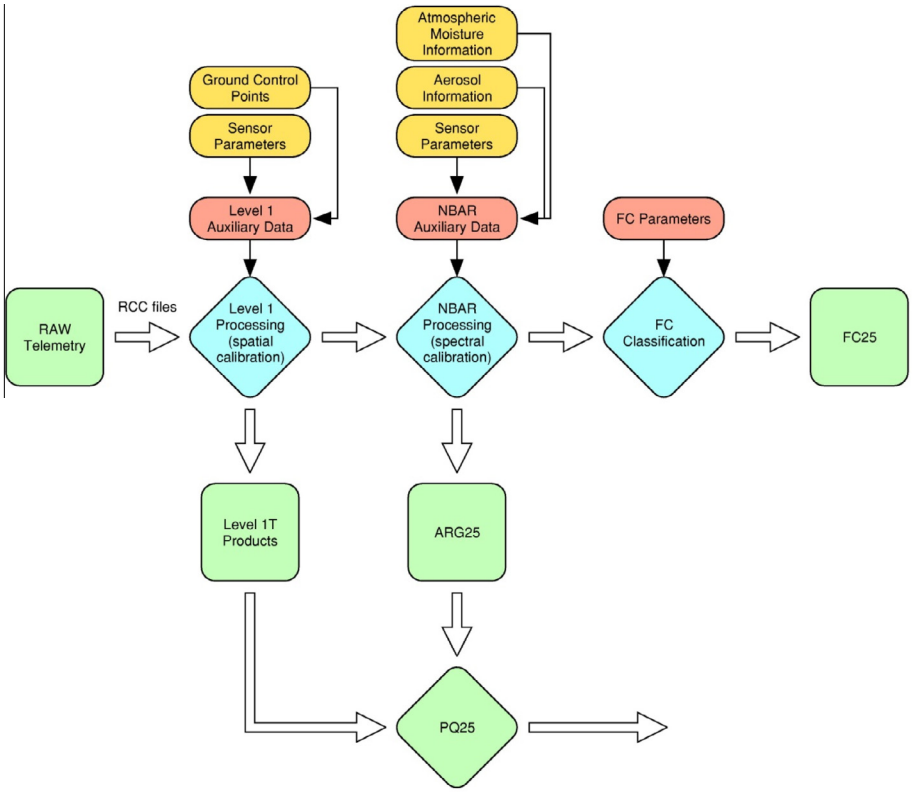

Fig. 4. High level workflow for Landsat data processing. RCC files refer to files containing raw telemetry data. L1T and L1G data products refer to Level 1 Standard Terrain Corrected Data and Level 1 Systematic Terrain Corrected Data respectively. ARG25 refers to 25 m Australian Reflectance Grid products. PQ25 refers to the Pixel Quality Assessment Product and FC25 refers to the fractional cover product derived from the ARG25.

Acknowledgements

The authors acknowledge the financial contribution of the Commonwealth Department of Industry's Australian Space Research Program to support the delivery of the ULA Project. The outcomes of the ULA Project would not have been possible without the dedicated and consistent effort from all project partners.

The research and development conducted by the National Earth & Marine Observations (NEMO) Group at Geoscience Australia under the ULA Project has provided a solid scientific basis that underpins the ARG25 and Pixel Quality Assessment Products and has enabled these data products to be transformed into High Performance Data Infrastructures that are supporting the development of advanced applications of the Australian Landsat Archive.

The implementation of FC25 was made possible by new scientific and technical capabilities, the collaborative framework established by the Terrestrial Ecosystem Research Network (TERN) through the National Collaborative Research Infrastructure Strategy (NCRIS), and the leadership and capabilities of Geoscience Australia and the Joint Remote Sensing Research Program.

The long term acquisition plan supported by the Landsat program (a joint initiative between the U.S. Geological Survey [USGS] and NASA); in combination with Geoscience Australia's longstanding investment in the on-ground infrastructure required to capture Landsat imagery has made the development of these data products possible.

The authors would like to thank the detailed and constructive comments provided by the external reviewers. Their suggestions have significantly improved this paper.

This paper has been published with the permission of the Chief Executive Officer, Geoscience Australia.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.grj.2015.02.010.

References

[1] Cohen WB, Goward SN. Landsat's role in ecological applications of remote sensing. BioScience 2004;54(6):535–45. http://dx.doi.org/10.1641/0006-3568(2004)054[0535:LRIEAO]2.0.CO;2.

[2] Wulder MA, White JC, Goward SN, Masek JR, Irons JR, et al. Landsat continuity: issues and opportunities for land cover monitoring. Remote Sens Environ 2008;112(3):955–69. http://dx.doi.org/10.1016/j.rse.2007.07.004.

[3] Woodcock CE, Allen R, Anderson M, Belward A, Bindschadler R, et al. Free access to landsat imagery. Science 2008;320:1011. http://dx.doi.org/10.1126/science.320.5879.1011a.

[4] Wulder MA, Masek JG, Cohen WB, Loveland TR, Woodcock CE. Opening the archive: how free data has enabled the science and monitoring promise of landsat. Remote Sens Environ 2012;122:2–10. http://dx.doi.org/10.1016/j.rse.2012.01.010.

[5] Australian Government. Unlocking the LANDSAT archive for future challenges; April 2013 [Online]. Available: http://www.space.gov.au/AustralianSpaceResearchProgram/ProjectFactsheetspage/Pages/UnlockingtheLANDSATArchiveforFutureChallenges.aspx [Accessed 30 January 2015].

[6] National Computational Infrastructure. Unlocking the Landsat Archive; 2013 [Online]. Available: http://nci.org.au/research/unlocking-the-landsat-archive/ [Accessed 30 January 2015].

[7] Purss MBJ, Lewis A, Frankish G, Chan TO, Evans B, Hurst L. Unlocking the landsat archive: enabling the future of earth observation science, In: 6th eResearch Australasia Conference, Sydney; 2012.

[8] US Geological Survey. Landsat Processing Details; 12 August 2014 [Online]. Available: http://landsat.usgs.gov/Landsat_Processing_Details.php [Accessed 30 January 2015].

[9] Li F, Jupp DL, Reddy S, Lymburner L, Mueller N, Tan P, et al. An evaluation of the use of atmospheric and BRDF correction to standardize landsat data. Sel Top Appl Earth Obs Remote Sens IEEE J 2010;3(3):257–70. http://dx.doi.org/10.1109/JSTARS.2010.2042281.

[10] Li F, Jupp DL, Thankappan M, Paget M, Lewis A, Held A. The variability of satellite derived surface BRDF shape over Australia from 2001 to 2011. In:

International Geoscience and Remote Sensing Symposium (IGARSS) – IEEE International; 2013. p. 255–8. http://dx.doi.org/10.1109/IGARSS.2013.6721140.

[11] Li F, Jupp DL, Thankappan M, Lymburner L, Mueller N, Lewis A, et al. A physics-based atmospheric and BRDF correction for landsat data over mountainous terrain. Remote Sens Environ 2012;124:756–70. http://dx.doi.org/10.1016/j.rse.2012.06.018.

[12] Irish RR, Barker JL, Goward SN, Arvidson T. Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogramm Eng Remote Sens 2006;72:1179.

[13] Zhu Z, Woodcock CE. Object-based cloud and cloud shadow detection in landsat imagery. Remote Sens Environ 2012;118:83–94. http://dx.doi.org/10.1016/j.rse.2011.10.028.

[14] Sixsmith J, Oliver S, Lymburner L. A hybrid approach to automated landsat pixel quality. In: Geoscience and remote sensing symposium (IGARSS), 2013 IEEE international; 2013. p. 4146–9. http://dx.doi.org/10.1109/IGARSS.2013.6723746.

[15] Scarth P, Roder A, Schmidt M. Tracking grazing pressure and climate interaction – the role of landsat fractional cover in the time series analysis. In: Proceedings of the 15th Australasian remote sensing & photogrammetry conference; 2010.

[16] Muir J, Schmidt M, Tindall D, Trevithick R, Scarth P, Stewart J. Guidelines for field measurement of fractional ground cover: a technical handbook supporting the Australian collaborative land use and management program. Canberra: Queensland Department of Environment and Resource Management for the Australian Bureau of Agricultural and Resource Economics and Sciences; 2011.

[17] ANZLIC. National Nested Grid (NNG) specification guideline. Canberra: Australian and New Zealand Land Information Council; 2012.

[18] Chander G, Markham BL, Helder DL. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sens Environ 2009;113(5):893–903. http://dx.doi.org/10.1016/j.rse.2009.01.007.

[19] Roy DP, Ju J, Mbow C, Frost P, Loveland T. Accessing free landsat data via the internet: Africa's challenge. Remote Sens Lett 2010;1(2):111–7. http://dx.doi.org/10.1080/01431160903486693.

[20] Lewis A, Lymburner L, Purss MBJ, Evans B, Ip A, et al. New 'data cube' approach to realize the potential of earth observations from space, in preparation.

[21] Purss MBJ, Lewis A, Ip A, Lymburner L, Oliver S, et al. The Australian geoscience data cube. In: Morain S, editor. Manual for remote sensing. 4th ed. American Society for Photogrammetry and Remote Sensing (ASPRS); 2015.

[22] Li F, Jupp DLB, Lymburner L, Tan P, McIntyre A, et al. Characteristics of MODIS BRDF shape and its relationship with land cover classes in Australia, In: 20th international congress on modelling and simulation, Adelaide; 2013.

[23] Guerschman JP, Hill MJ, Renzullo LJ, Barrett DJ, Marks AS, et al. Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors. Remote Sens Environ 2009;113(5):928–45. http://dx.doi.org/10.1016/j.rse.2009.01.006.

[24] Veraverbeke S, Somers B, Gitas I, Katagis T, Polychronaki A, Goossens R. Spectral mixture analysis to assess post-fire vegetation regeneration using Landsat Thematic Mapper imagery: accounting for soil brightness variation. Int J Appl Earth Obs Geoinf 2011;14(1):1–11. http://dx.doi.org/10.1016/j.jag.2011.08.004.

[25] Purss MBJ, Lewis A, Edberg R, Ip A, Sixsmith J, et al. Exploiting data intensive applications on high performance computers to unlock Australia's landsat archive. In: Geophysical Research Abstracts, vol. 15; 2013. p. EGU2013-8049.

[26] Lymburner L, McIntyre A, Li F, Ip A, Thankappan M, Sixsmith J. Creating multi-sensor time series using data from Landsat-5 TM and Landsat-7 ETM+ to characterise vegetation dynamics. In: IEEE international geoscience and remote sensing symposium, Melbourne, Australia; 2013. http://dx.doi.org/10.1109/IGARSS.2013.6721321.

[27] Elsevier Research Data Services. Geoscience data rescue award 2013 submissions – unlocking the landsat archive; 2013 [Online]. Available: http://researchdata.elsevier.com/datachallenge/submission/16 [Accessed 30 January 2015].

[28] Showstack R. Award program recognizes efforts to protect geoscience data. EOS Trans AGU 2014;95(1):2. http://dx.doi.org/10.1002/2014EO010002.