Australian Government
Geoscience Australia

Australian Geoscience Data Cube

CEOS WGISS-39

Simon Oliver
Jonathan Ross

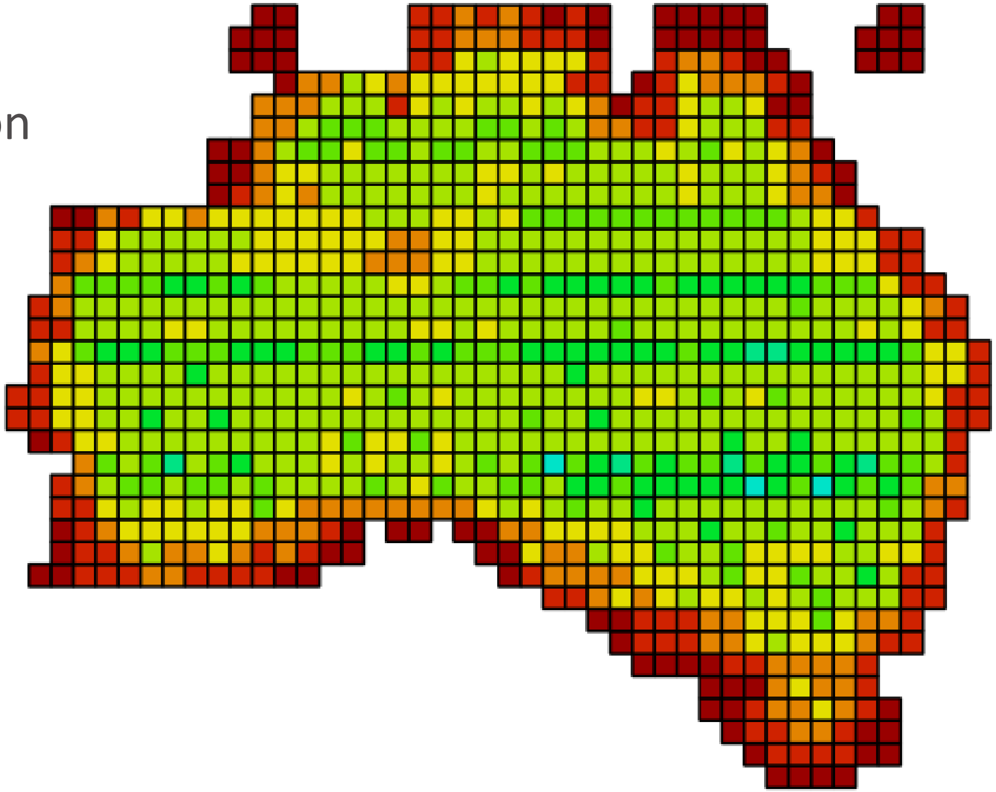
NCI

Australian Government
Geoscience Australia

CSIRO

Overview

- Geoscience Australia background and EO history
- Introduction to the AGDC: common analytical platform for EO data
- Example Applications
- EO Data Collection Management
- AGDC API overview and usage



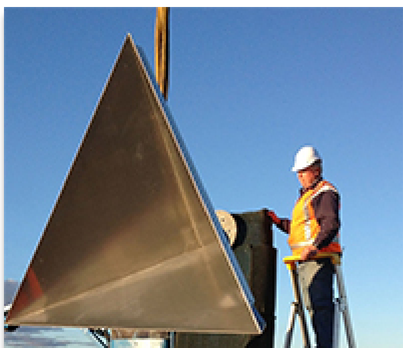
Australian Government
Geoscience Australia

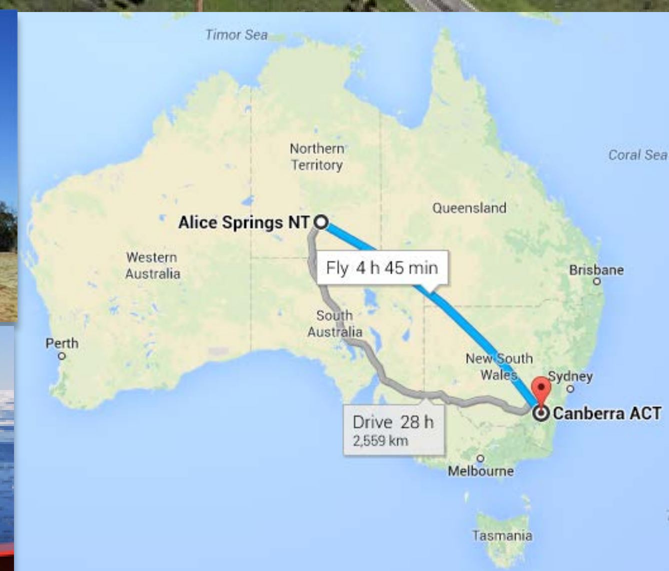
CSIRO

NCI

Geoscience Australia – background and EO history

Geoscience Australia applies geoscience to Australia's most important challenges by providing geoscience information, services and capability to the Australian Government, industry and stakeholders.



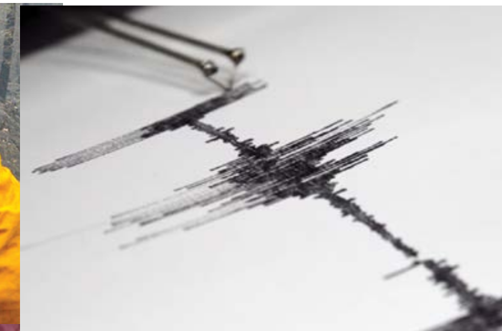


Australian Government
Geoscience Australia

CSIRO

NCI

Geoscience Australia - background and EO history



Strategic priorities:

1. Building Australia’s Resource Wealth

2. Ensuring Australia’s Community Safety

3. Managing Australia’s Marine Jurisdictions

4. Securing Australia’s Water Resources

5. Providing Fundamental Geographic Information

6. Maintaining Geoscience Knowledge and Capability

Geoscience Australia – background and EO history

Energy Division	Community Safety and Earth Monitoring Division	Environmental Geoscience Division	Corporate Services
Energy Systems	Community Safety	Groundwater	ICT Innovation and Services
Mineral Systems	Geodesy and Seismic Monitoring	National Earth and Marine Observations	
Resources Advice and Promotion	Observatories and Science Support	National Location Information	

Geoscience Australia is a publicly funded Agency within the Australian Government Industry and Science portfolio

~AUD $130M budget for financial year 2013/14

Australian Government
Geoscience Australia

CSIRO

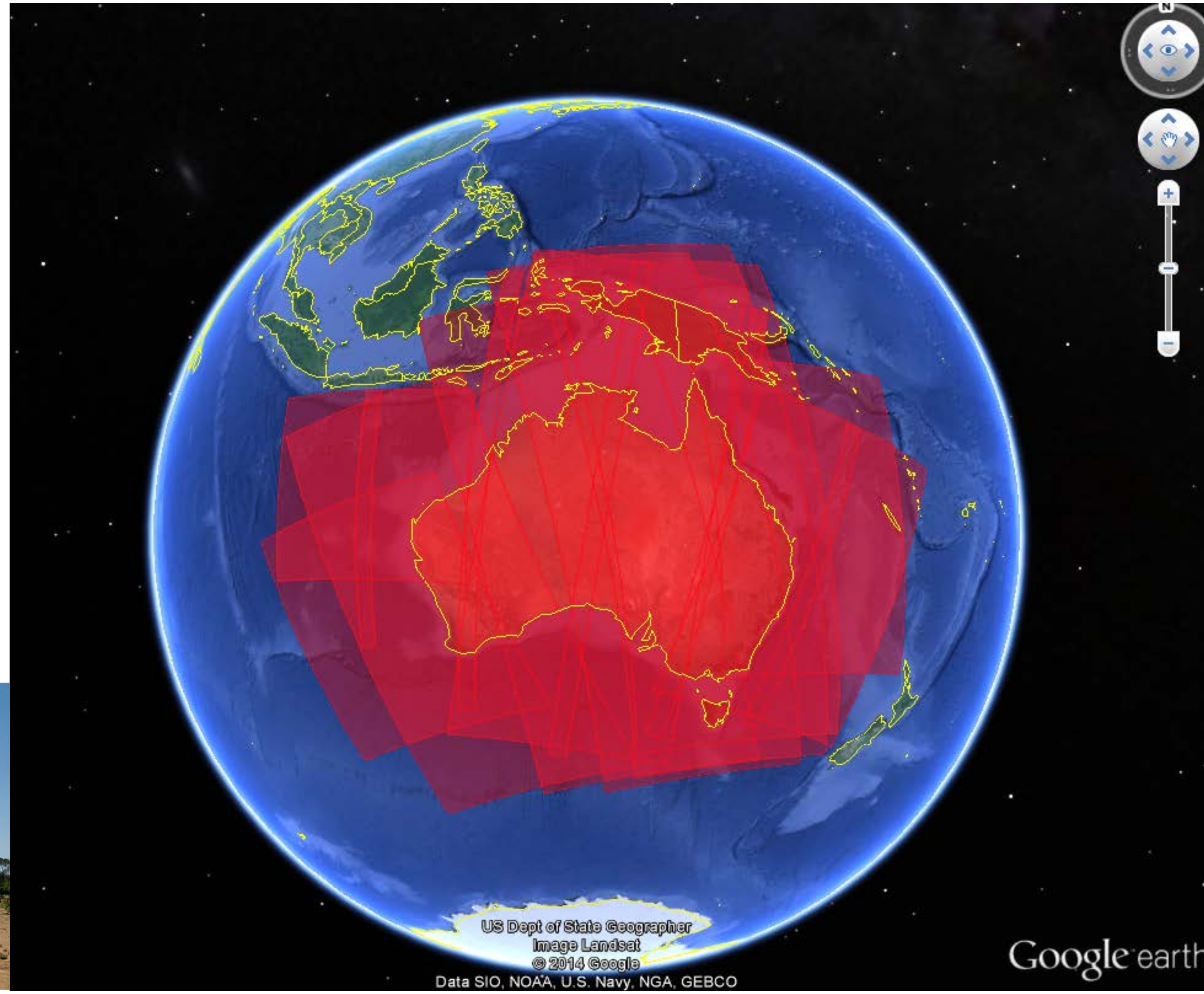
NCI

Relationships

- Historically strong relationship with US Government on Landsat mission support
- Support for ALOS, JERS
- Membership and active participant in Landsat Science Team
- NASA Systems Engineering Office for KenyaCube in support of GFOI/GEOGLAM
- Developing a Memorandum of Understanding with the European Commission / ESA
- Seeking to engage more closely with ESA regarding Copernicus

A typical day of data acquisition

- Landsat7
- Landsat8
- Terra MODIS
- Aqua MODIS
- Suomi NPP
- NOAA AVHRR



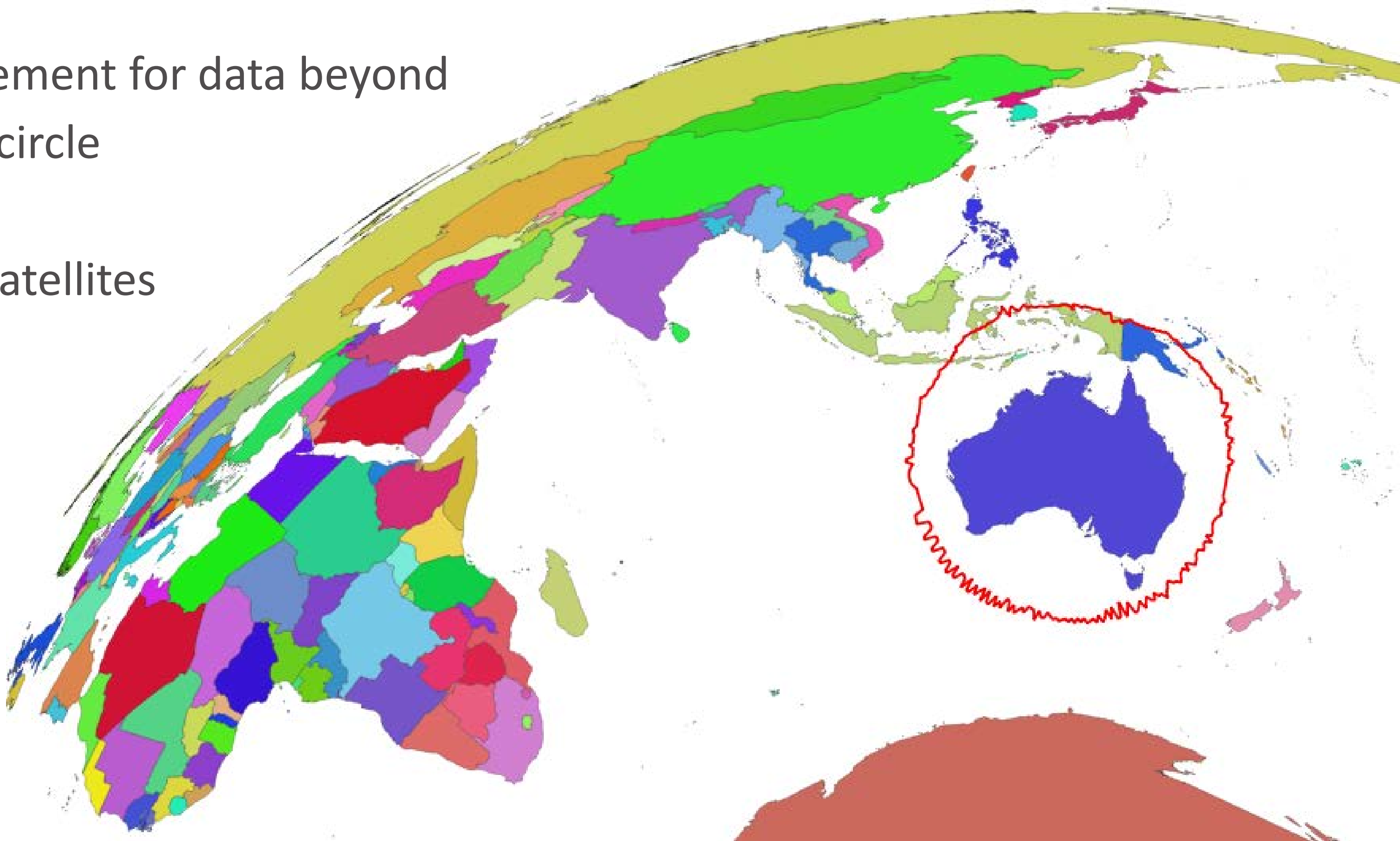
From direct reception to internet bulk transfer

Downlink at Alice Springs reception facility

High volume data transfers via the internet

Growing requirement for data beyond our acquisition circle

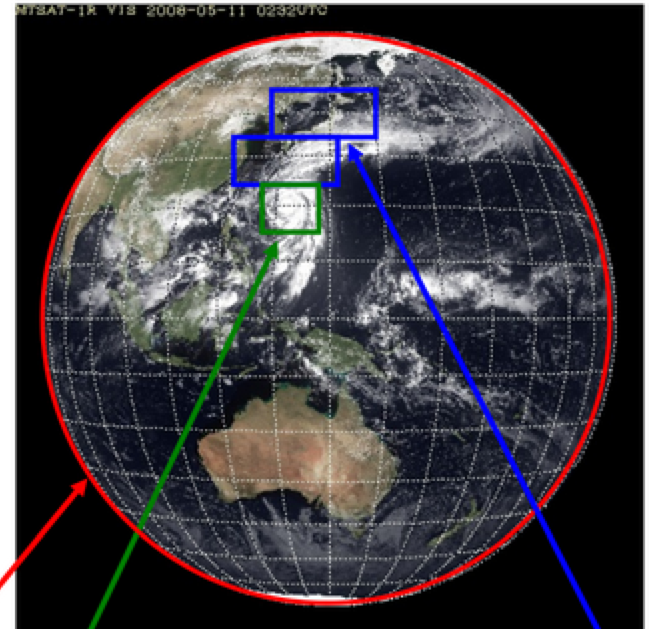
Geostationary satellites



Himawari-8: Specification of Observation

Bands of Himawari-8/9

Band	Wavelength [μm]	Spatial Resolution	
1	0.47	1 km	RGB Composited True Color Image
2	0.51	1 km	
3	0.64	0.5 km	
4	0.86	1 km	
5	1.6	2 km	
6	2.3	2 km	
7	3.9	2 km	
8	6.2	2 km	Water Vapor
9	6.9	2 km	
10	7.3	2 km	
11	8.6	2 km	SO_2
12	9.6	2 km	O_3
13	10.4	2 km	Atmospheric Windows
14	11.2	2 km	
15	12.4	2 km	
16	13.3	2 km	CO_2



Full disk
Interval: **10 minutes** (6 times per hour)

Japan Area
Interval: **2.5 minutes** (4 times in 10 minutes)
Dimension: EW x NS: 2000 x 1000 km x 2

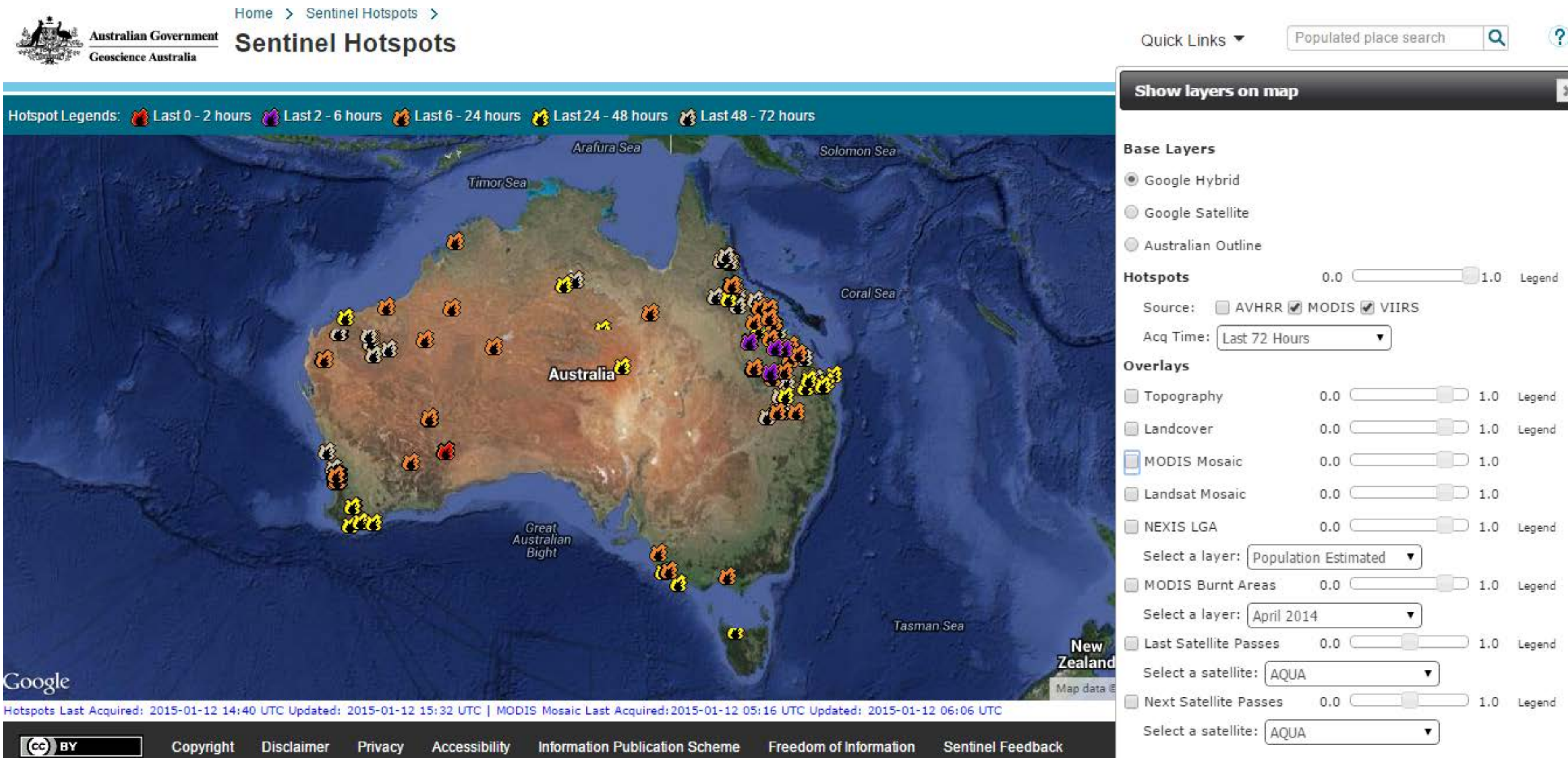
Target Area
Interval: **2.5 minutes** (4 times in 10 minutes)
Dimension: EW x NS: 1000 x 1000 km

Number of Bands: 5 → 16

Interval: 30/60 min. → 10 min.

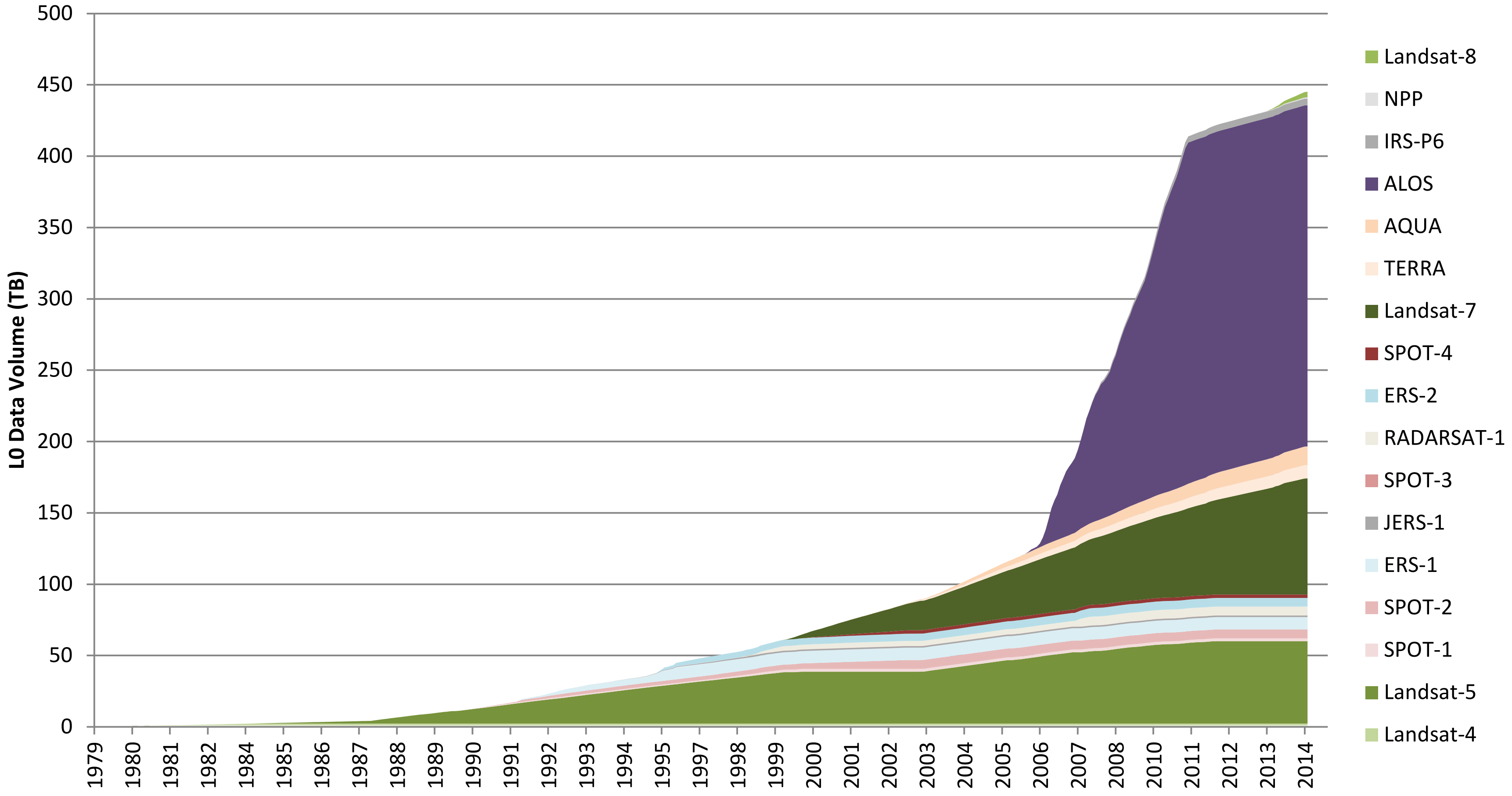
http://severe.worldweather.wmo.int/TCFW/JMAworkshop/4-3.Himawari8-9_YIzumikawa.pdf

From Detection to Situational Awareness



http://sentinel.ga.gov.au/

Satellite Earth Observation Data Holdings at Geoscience Australia 1979 – 2014 (L0)

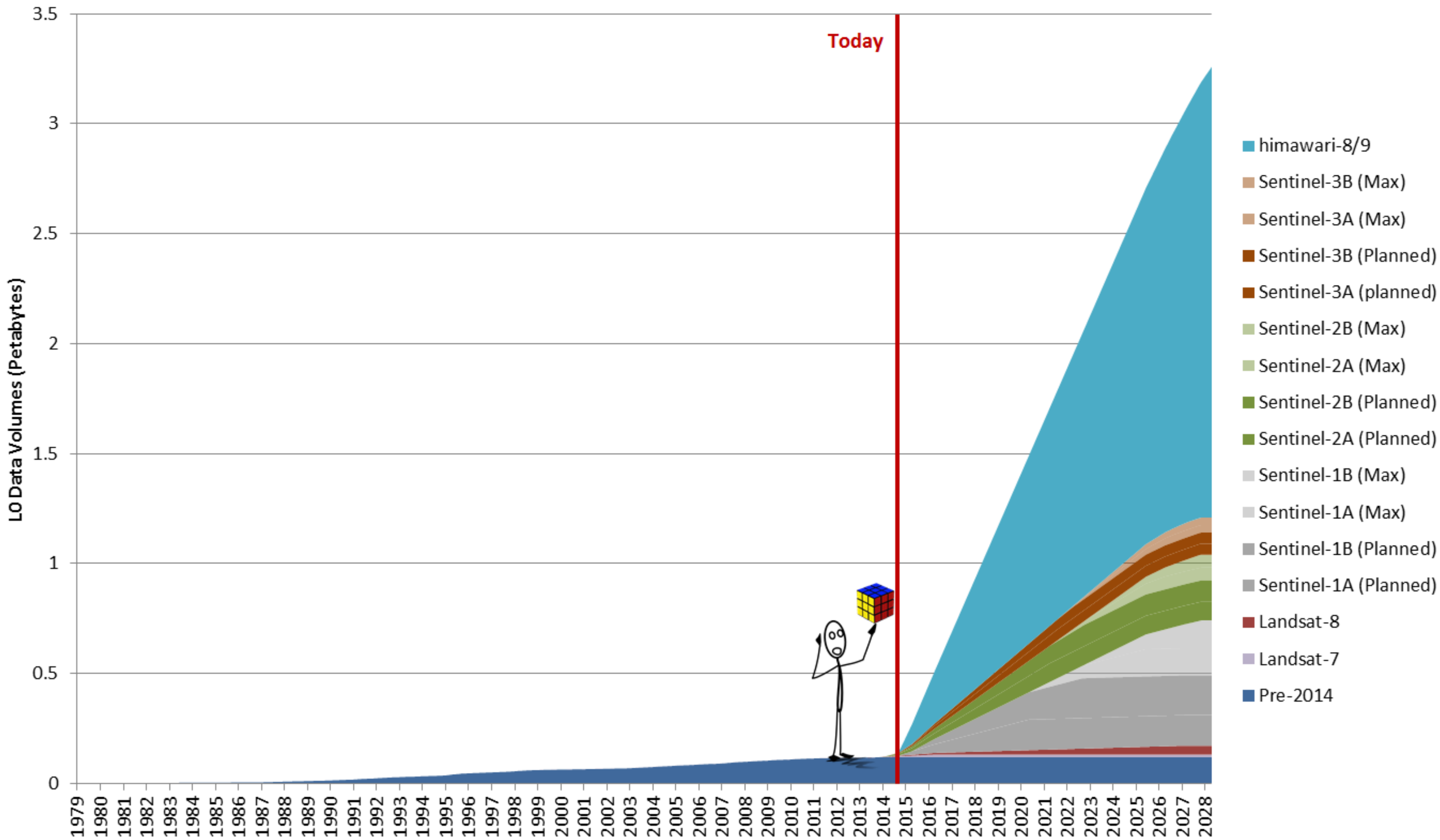


Australian Government
Geoscience Australia

CSIRO

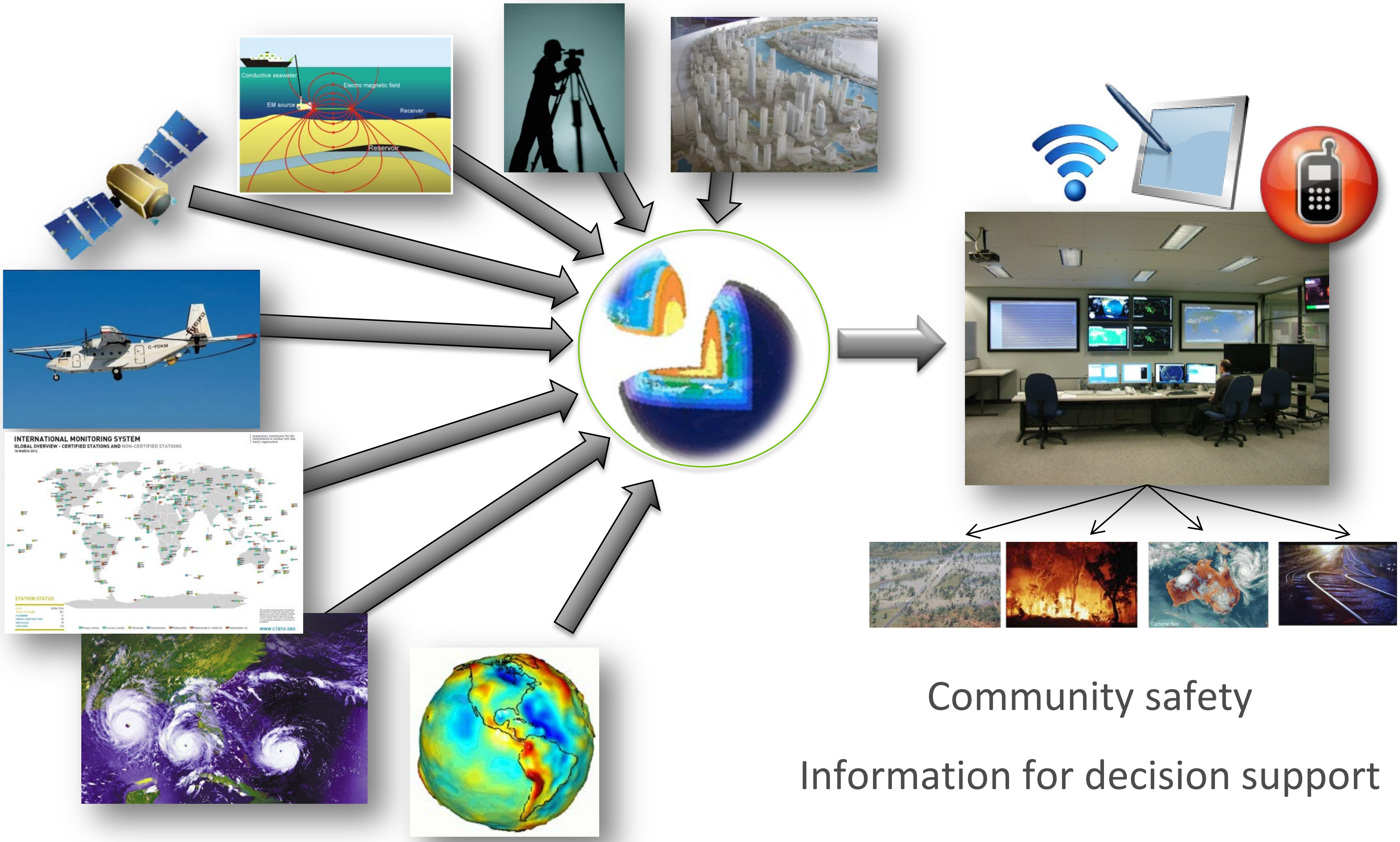
NCI

In the next decade



Australian Government
Geoscience Australia
CSIRO
NCI

The growing expectations of users



Community safety

Information for decision support

Australian Government
Geoscience Australia

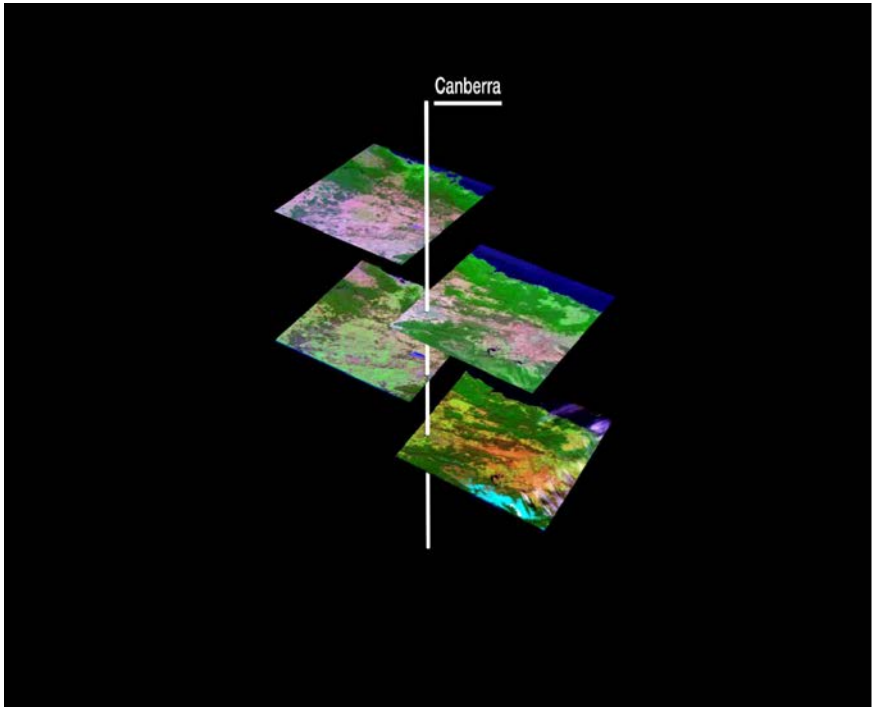
CSIRO

NCI

The challenge

Data collection is dynamic: growing in time, and also subject to modification (existing data) and insertion (new data). The challenge is to enable:

- Attribution of exact observation time for key applications e.g. tides for shallow-water bathymetry, bare earth
- Analysis of each observation in the time-series
- Reliable comparison of observations over long periods of time, e.g. change detection, pattern analysis
- Iteration and refinement of processes at continental scale
- Rapid generation of results



Unlocking the Landsat Archive

A Space Science and Innovation project that received **AUD $3,472,965 (3 years)** funding through the Australian Space Research Program.

- **Lockheed Martin** Australia Pty Ltd (LMA)
- Australian National University **National Computational Infrastructure** (NCI)
- **Geoscience Australia** (GA)
- Victorian Partnership for Advanced Computing Ltd (VPAC)
- Cooperative Research Centre for Spatial Information (CRCSI)

Outcomes

- Migrated Australia's Landsat archive from tape to spinning disk
- Developed processing routines for automated calibration of data
- Prototype development of the Australian Geoscience Data Cube

Australia Geoscience Data Cube

NCI

Australian Government
Geoscience Australia

CSIRO

High Performance Computing

- Raijin @ National Computational Infrastructure
- **AUD $50M** to buy – **AUD $12M**/year to operate
- **57,472 cores** (Intel Xeon Sandy Bridge technology, 2.6 GHz) in 3592 compute nodes;
- 160 TBytes (approx.) of main memory;
- **10 PBytes (approx.) of usable fast filesystem (for short-term scratch space).**

37	Research Institute for Information Technology, Kyushu University Japan	**QUARTETTO** - HA8000-tc HT210/PRIMERGY CX400 Cluster, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, NVIDIA K20/K20x, Xeon Phi 5110P Hitachi/Fujitsu
38	National Computational Infrastructure, Australian National University Australia	Fujitsu PRIMERGY CX250 S1, Xeon E5-2670 8C 2.600GHz, Infiniband FDR Fujitsu
39	Purdue University United States	**Conte** - Cluster Platform SL250s Gen8, Xeon E5-2670 8C 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P Hewlett-Packard

*http://top500.org/

Australian Government
Geoscience Australia
CSIRO
NCI

National Computational Infrastructure

NCI operates as a formal collaboration of a number of research institutions. The major partners are:

- Australian National University (ANU)
- Commonwealth Scientific and Industrial Research Organisation (CSIRO)
- Australian Bureau of Meteorology (BoM)
- Geoscience Australia (GA) 4% share AUD $0.5M / 3 years

NCI PROVIDING AUSTRALIAN RESEARCHERS WITH WORLD-CLASS HIGH-END COMPUTING SERVICES

Australian Government
Geoscience Australia

CSIRO

NCI

Funding supporting data infrastructure

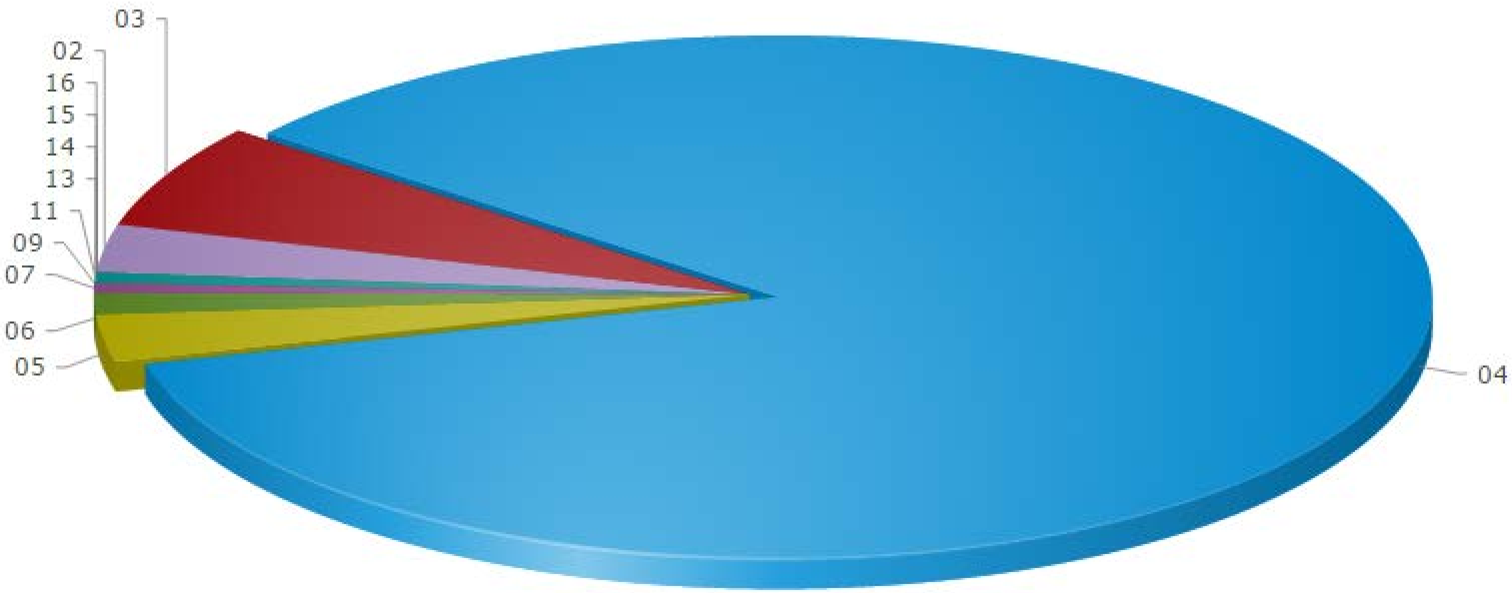
Australian Government Initiatives helping build the foundations for EO data exploitation

RDSI **AUD $50M** to enhance data centre development and support retention and integration of nationally significant data assets into the national collaboration and data fabric.

NCRIS
National Collaborative Research Infrastructure Scheme

NCI PROVIDING AUSTRALIAN RESEARCHERS WITH WORLD-CLASS HIGH-END COMPUTING SERVICES

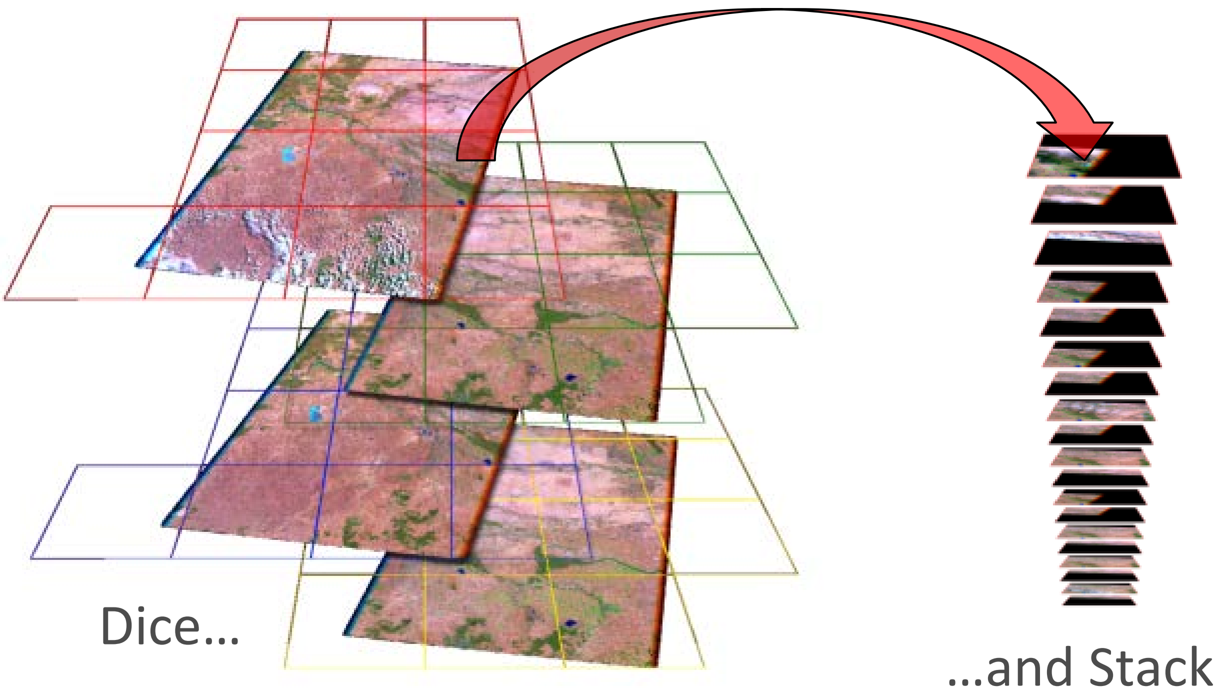
RDSI Research Data Storage Infrastructure



- 02 Physical Sciences - 2.92%
- 03 Chemical Sciences - 6.44%
- 04 Earth Sciences - 85.1%
- 05 Environmental Sciences - 2.92%
- 06 Biological Sciences - 1.34%
- 07 Agricultural and Veterinary Sciences - 0.54%
- 09 Engineering - 0.05%
- 11 Medical and Health Sciences - 0.65%
- 13 Education - 0.01%
- 14 Economics - 0.01%
- 15 Commerce, Management, Tourism and Services - 0.01%
- 16 Studies in Human Society - 0.01%

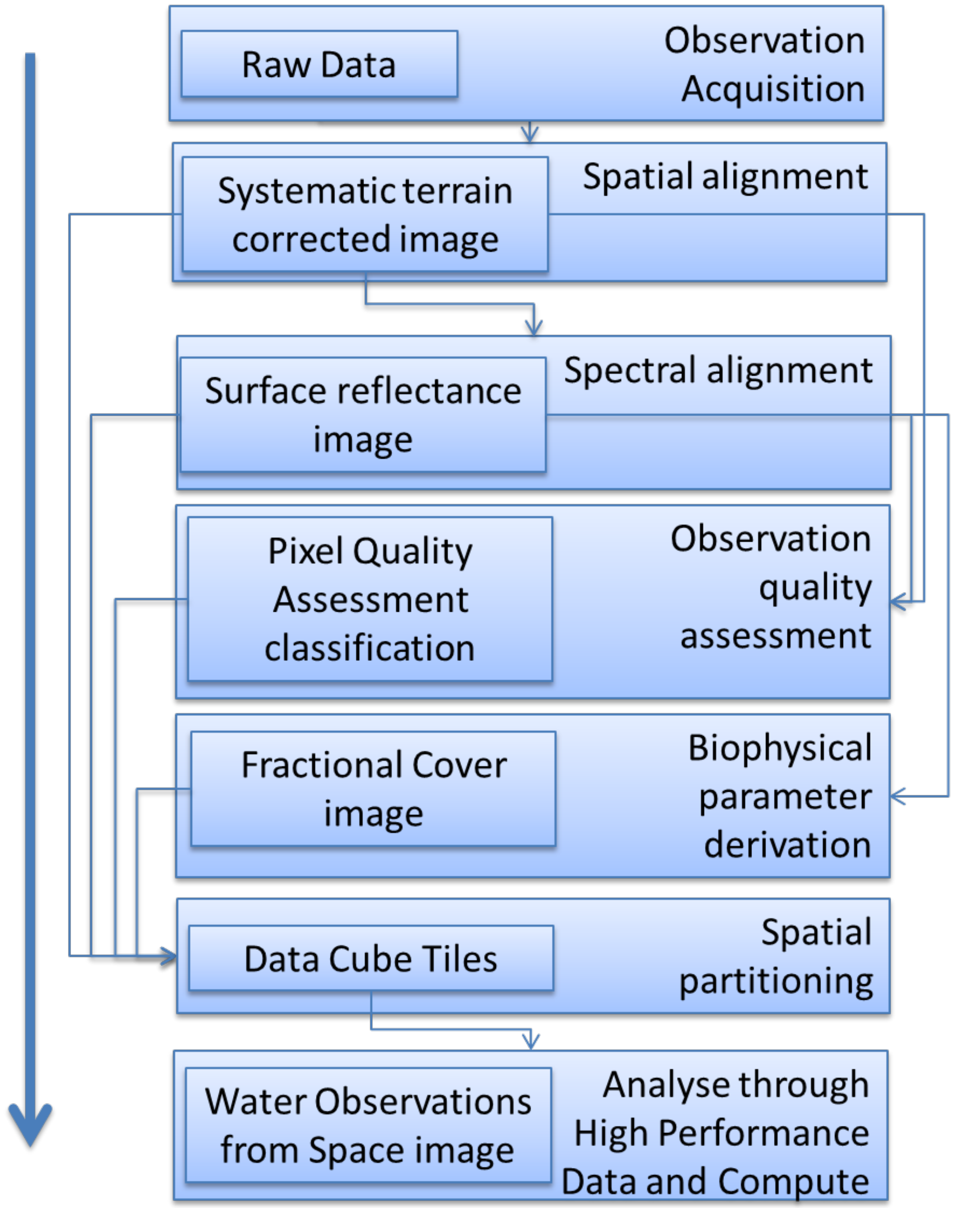
Simplified data structures

- The AGDC arranges 2D (spatial) data temporally and spatially to allow flexible but reasonably efficient large-scale analysis.
- “Dice’n’Stack” method used to subdivide the data into spatially-regular, time-stamped, band-aggregated tiles which can be managed as dense temporal stacks.



Australian Government Geoscience Australia CSIRO NCI

Robust and highly iterable processes



Raw Computer Compatible format satellite telemetry data
Courtesy of United States Geological Survey Organisation

Landsat Product Generation System – Systematic Terrain Correction (ortho-correction)
Courtesy of United States Geological Survey Organisation

Nadir Bi-directional Reflectance Distribution Function Adjusted Reflectance [1]

Pixel Quality Assessment – 16 bit binary tests representing results of observation quality assessment [3]

Fractional Cover [3] probability assessment:
- bare ground fraction
- green vegetation fraction
- non-green vegetation
- Mask Layer

Courtesy of Joint Remote Sensing Research Program

Tiling and aggregation of image layers to simplify access to Data Cube components

1. Li, F., Jupp, D. L., Reddy, S., Lymburner, L., Mueller, N., Tan, P., & Islam, A. (2010). An evaluation of the use of atmospheric and BRDF correction to standardize Landsat data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, *3*(3), 257-270.

2. *Scarth, P, Roder, A and Schmidt, M 2010, 'Tracking grazing pressure and climate interaction – the role of Landsat fractional cover in time series analysis', Proceedings of the 15th Australasian Remote Sensing & Photogrammetry Conference, viewed 4 January 2011*

3. Sixsmith, J., Oliver, S., & Lymburner, L. (2013, July). National Earth Observation Group, Geoscience Australia, GPO Box 378, ACT 2601, Australia. In *Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International* (pp. 4146-4149). IEEE.

Australian Government
Geoscience Australia

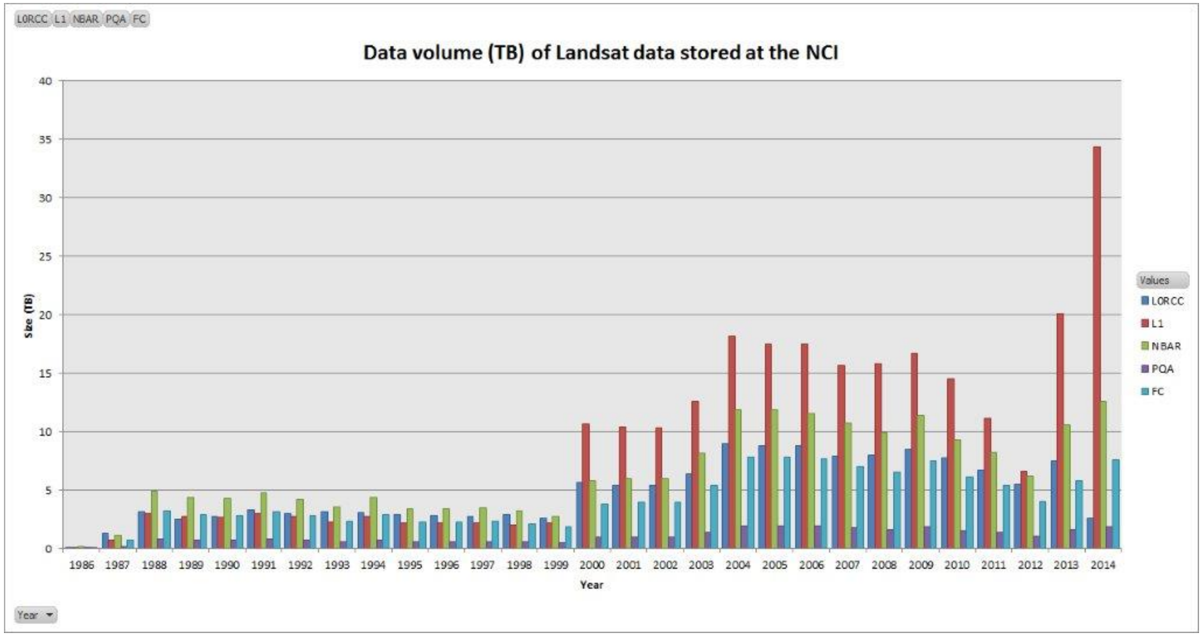
CSIRO

NCI

Current AGDC contents - Landsat

27 years of Landsat data (1987-2014) processed so far*:

- 20,500 passes in 301,400 acquisitions
- 857,000 available datasets (all processing levels)
- 93 x 10^{12} pixels in all available datasets
- ~1200 observations for some areas
- **~0.75PB**



* Figures as at 30/9/14 rounded to nearest hundred

Support for other observation platforms

Level 1 Topographic (ORTHO)

1. LS5-B60 – Thermal Infrared
2. LS7-B61 – Thermal Infrared Low Gain
3. LS7-B62 – Thermal Infrared High Gain

ARG-25 (NBAR)

1. LS5/7-B10 – Visible Blue
2. LS5/7-B20 – Visible Green
3. LS5/7-B30 – Visible Red
4. LS5/7-B40 – Near Infrared
5. LS5/7-B50 – Middle Infrared 1
6. LS5/7-B70 – Middle Infrared 2

Pixel Quality (PQA)*

1. PQ – Bit-array of PQ tests

Fractional Cover (FC)**

1. Photosynthetic Veg. (PV)
2. Non-Photosynthetic Veg. (NPV)
3. Bare Soil (BS)
4. Un-mixing Error (UE)

Digital Elevation Model

(CC-by 1" DEM)

1. DEM - Bare-earth DEM
2. DEM-S - bare-earth DEM, adaptively smoothed
3. DEM–H – hydrologically enforced

MODIS

1. MOD09 – surface reflectance
2. MOD43 – NBAR corrected

ASTER

1. Mineral products

AGRI

AVHRR

MERIS

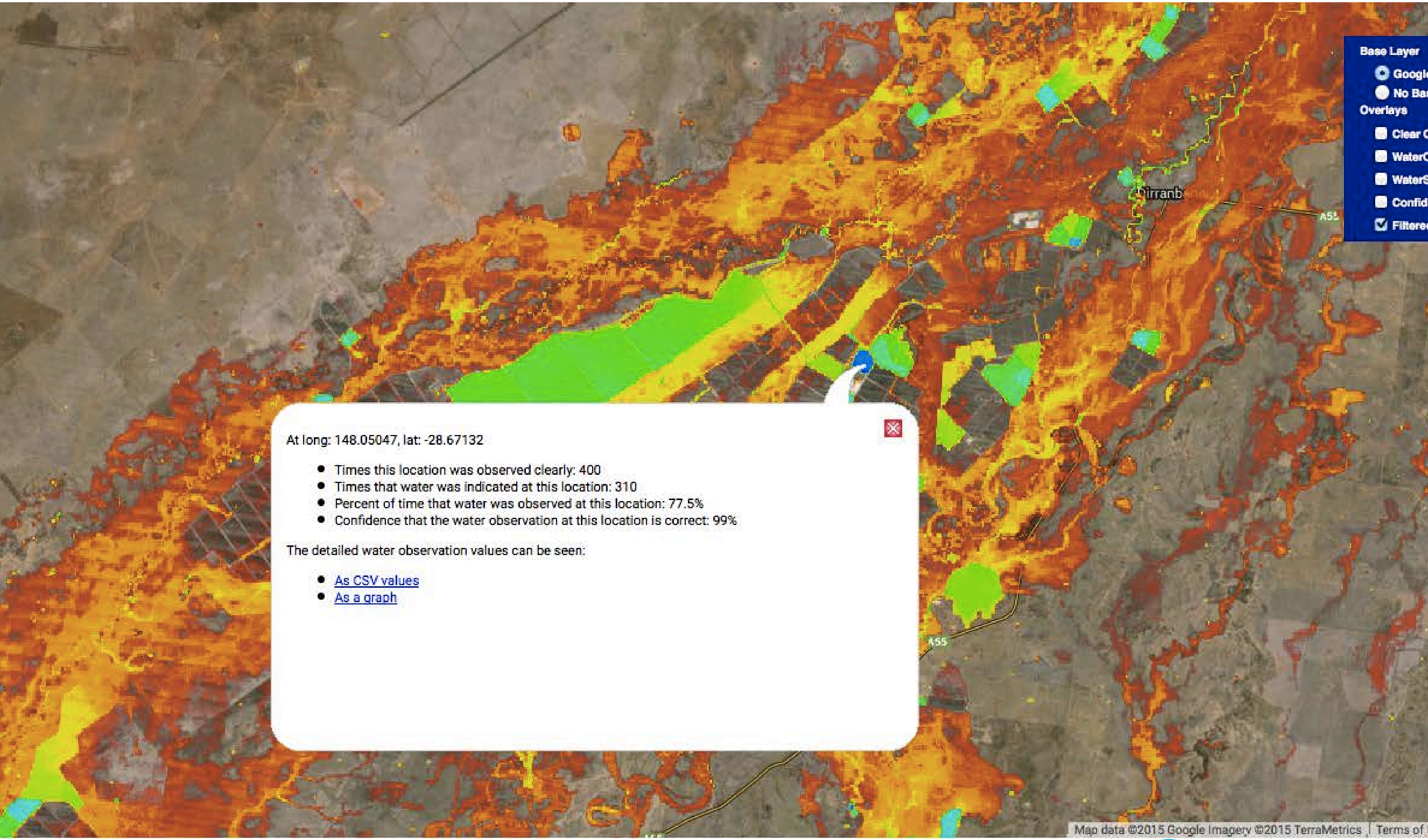
…

* PQA Geoscience Australia
** JRSRP

Example Applications – Water Observations from Space

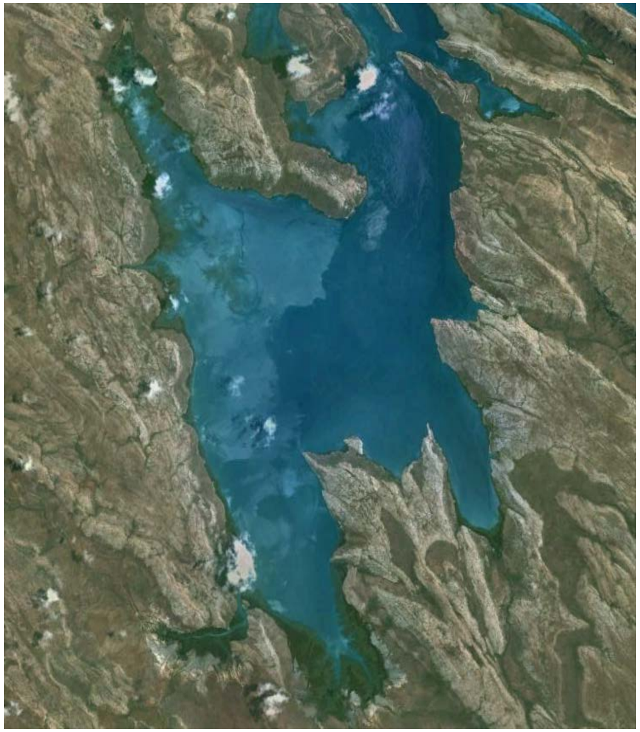
Australian Government
Geoscience Australia

Example Applications – Water Observations from Space



Australian Government
Geoscience Australia
CSIRO
NCI

Example Applications
- Using tidal models to map tidal extents



Tidal Range of >10m

Tidal Zone Extent

Can be attributed with offsets of LAT to lowest observed tide and HAT to highest observed

Tidal Zone Morphology

Fraction of water observations over the time series. Can we attribute this with depths?

Australian Government
Geoscience Australia

CSIRO

NCI

Example Applications
- National Fractional Cover Time Series
Joint Remote Sensing Research Program

The University of Queensland Australia

TERN Terrestrial Ecosystem Research Network

Queensland Government

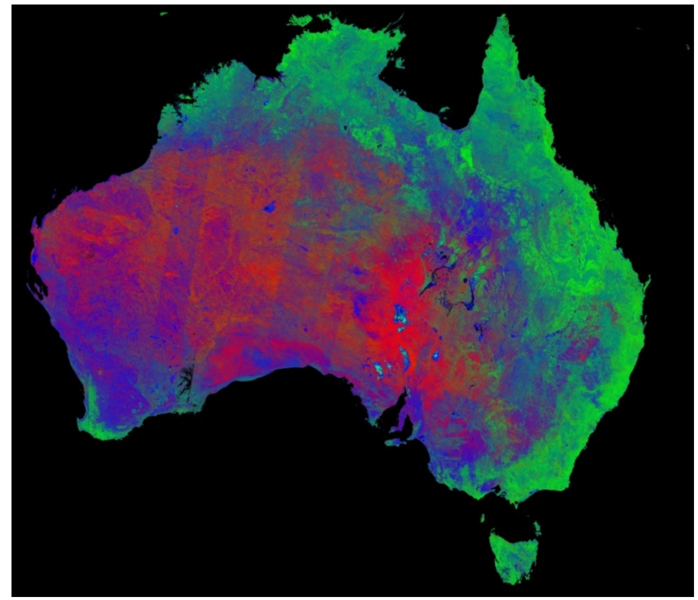
Fractional cover uses a constrained un-mixing model with end-members derived from field sampling.

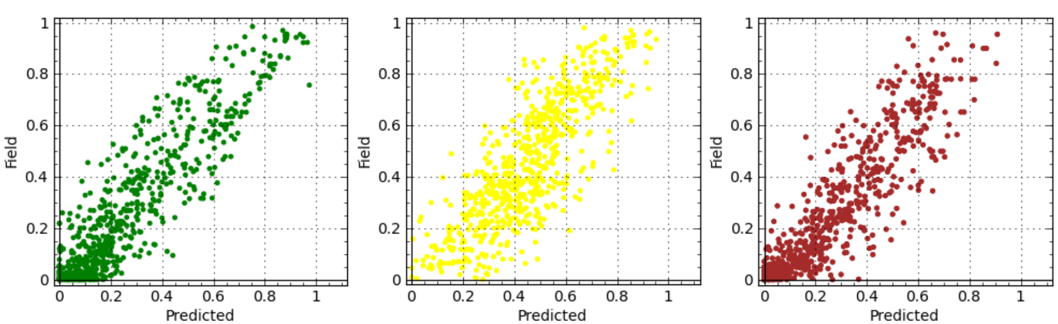
Creates an image with the percentage of bare, green and non-green fractions

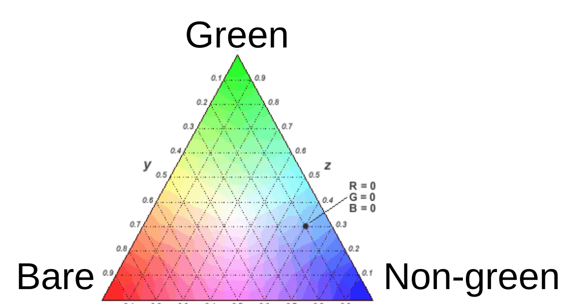
Over 1100 field sites collected using consistent, nationally agreed protocol

Captures cover dynamics at 25m* resolution

Applied by Geoscience Australia nationally using Australian Space Research Program funds



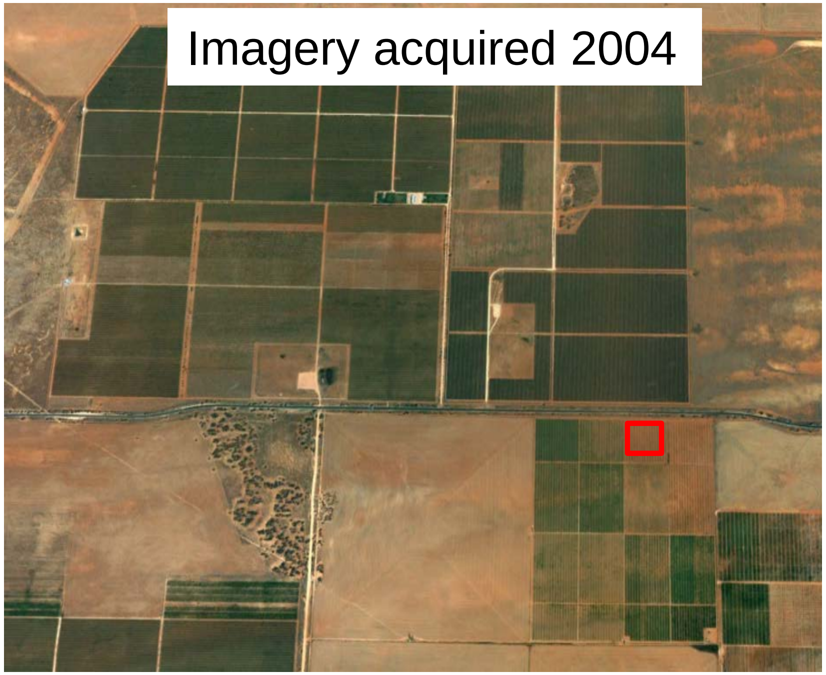


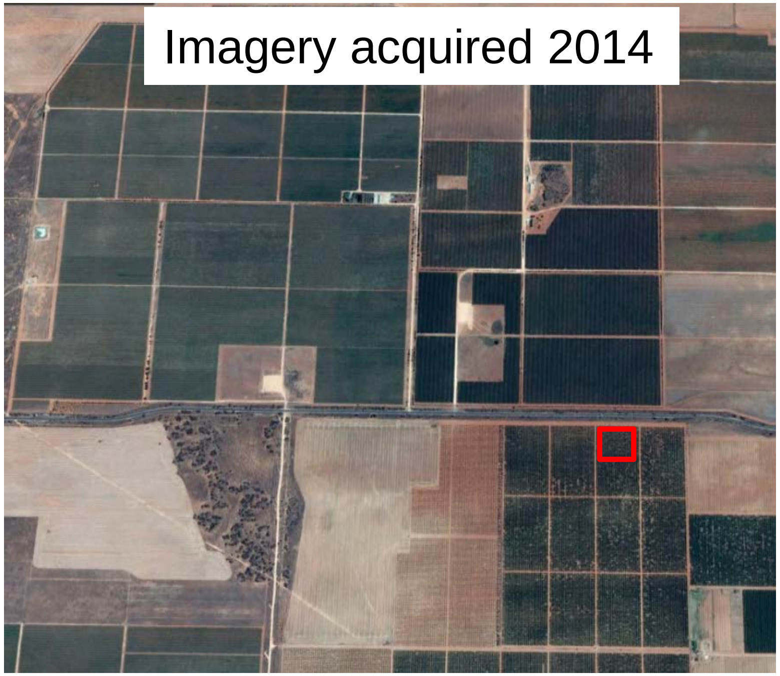


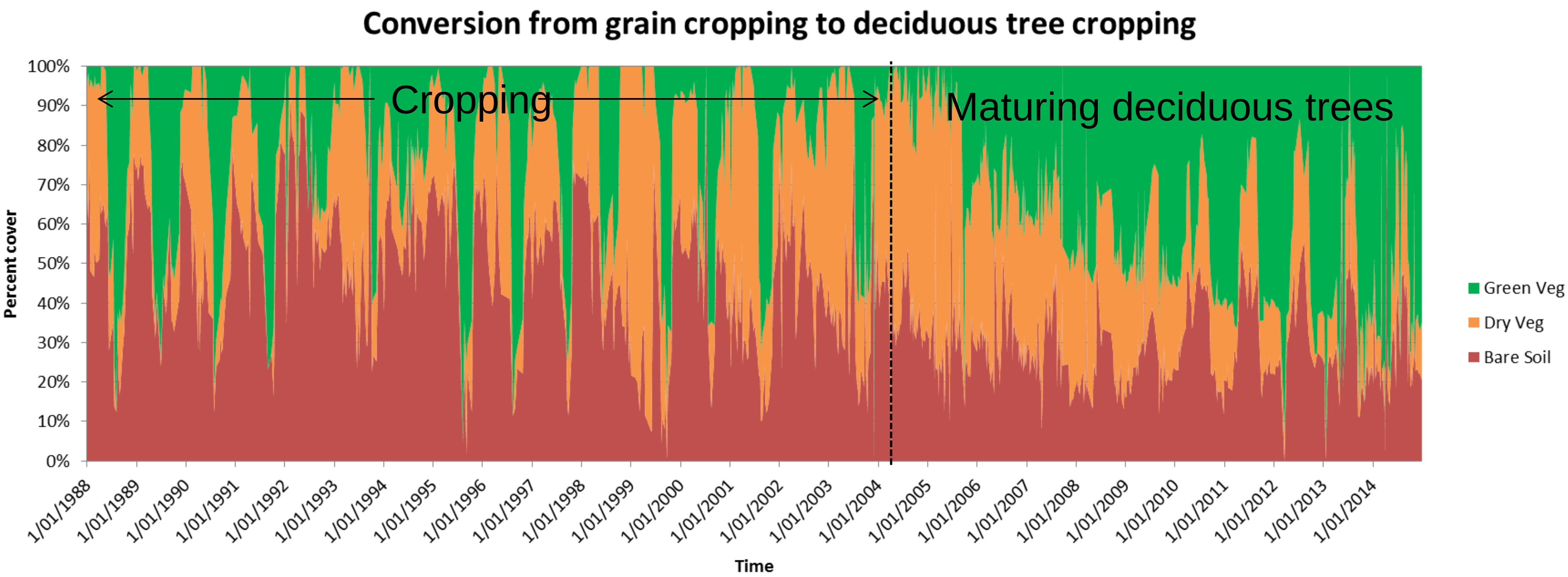
Australian Government Geoscience Australia

CSIRO

NCI

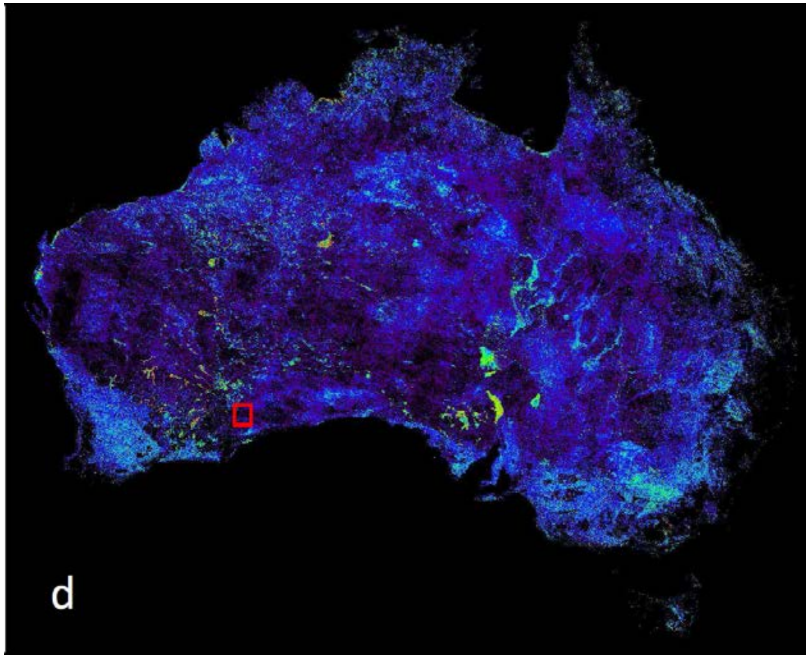

Imagery acquired 2004


Imagery acquired 2014


Conversion from grain cropping to deciduous tree cropping
Cropping
Maturing deciduous trees
Percent cover
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1/01/1988
1/01/1989
1/01/1990
1/01/1991
1/01/1992
1/01/1993
1/01/1994
1/01/1995
1/01/1996
1/01/1997
1/01/1998
1/01/1999
1/01/2000
1/01/2001
1/01/2002
1/01/2003
1/01/2004
1/01/2005
1/01/2006
1/01/2007
1/01/2008
1/01/2009
1/01/2010
1/01/2011
1/01/2012
1/01/2013
1/01/2014
Time
Green Veg
Dry Veg
Bare Soil

AGDC application themes to be supported

1. Water
   a) National Flood Risk Information
   b) Inland Water Detection
   c) Shallow water bathymetry/intertidal
2. Vegetation
   a) Condition Assessment
   b) Carbon Accounting
   c) Crop mapping & primary productivity
3. Data Fusion
   a) Landsat and MODIS Blending
4. Cal/Val site identification (detecting stable spectral response)
5. Geology
   a) Detecting bare earth to enhance mineral mapping



Australian Government
Geoscience Australia

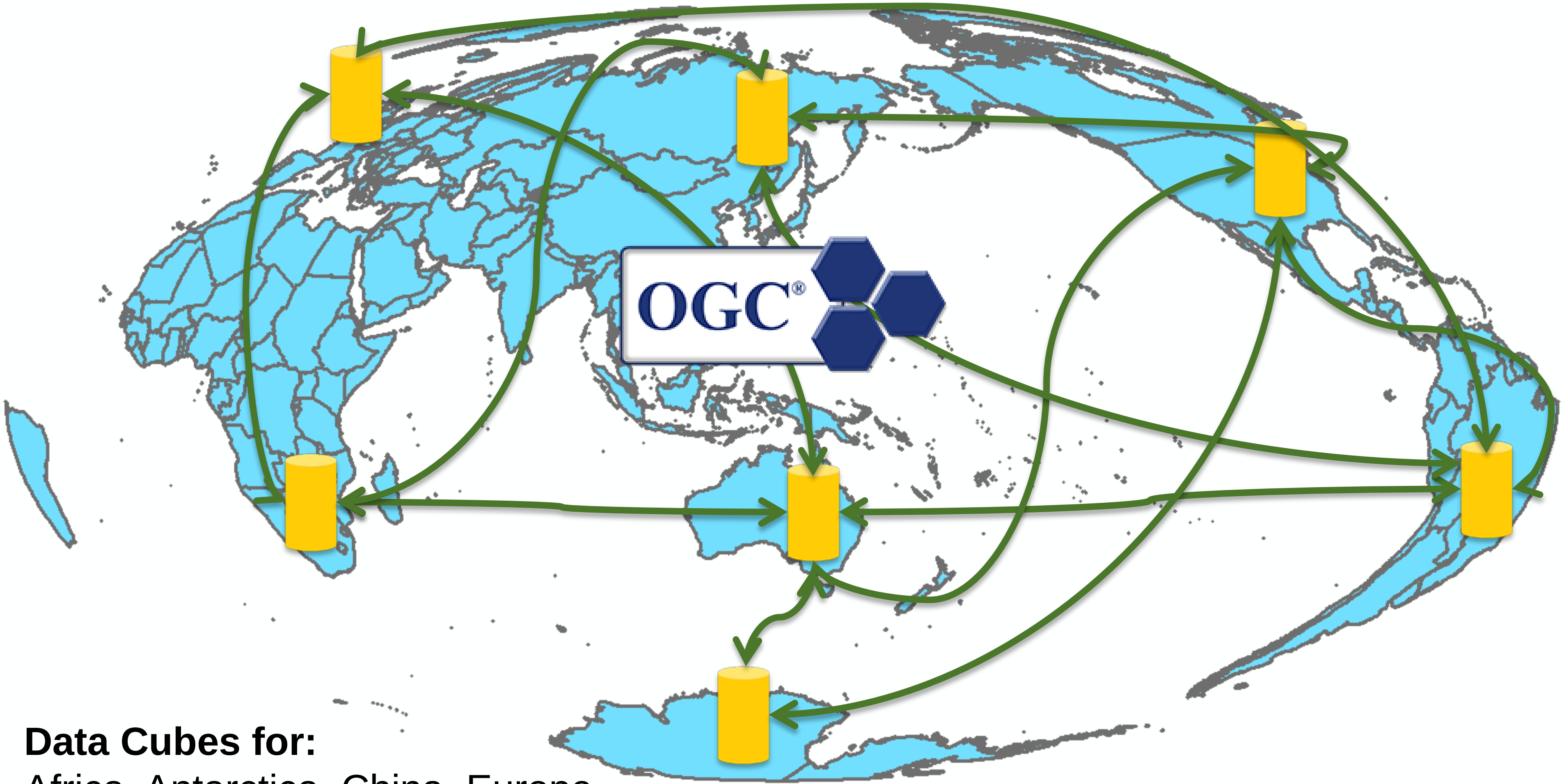
CSIRO

NCI

AGDC Contents

- **AGDC database** – provides indexing and filtering capability to enable attribute-based tile selection
- **AGDC API** – facilitates algorithm construction
- Written in **Python** and based on the open source Geospatial Data Abstraction Library / GDAL esp. Virtual Raster Transforms
- Data grid specification based on the ANZLIC National Nested Grid Specification Guide – **OGC DGGS SWG**

Discrete Global Grid System



Data Cubes for:
Africa, Antarctica, China, Europe, North America, …

Simplifying AGDC production tasks

- Luigi enables construction of **complex pipelines of long-running batch jobs** by handling dependency resolution, workflow management, visualization etc.

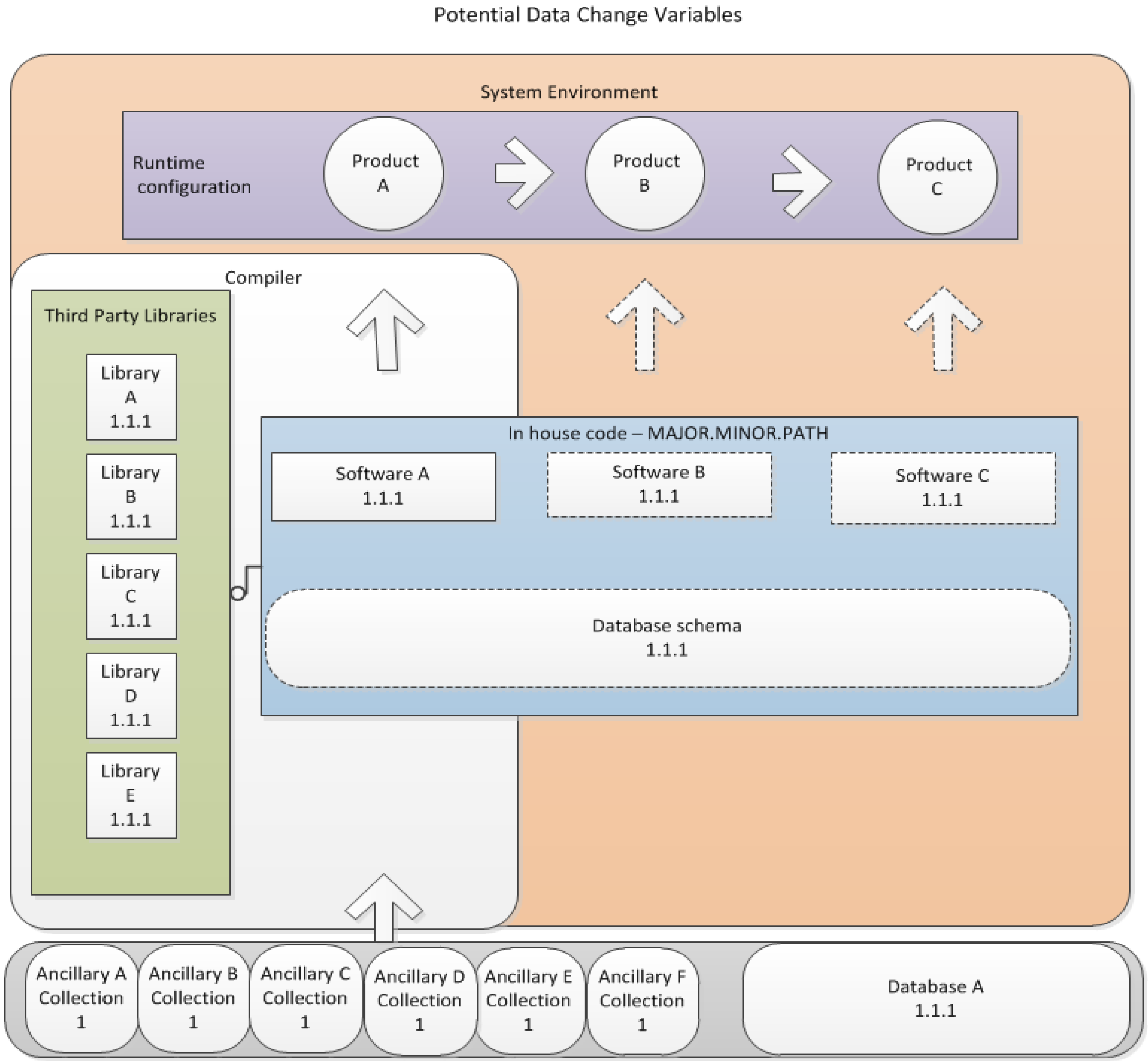
- Conceptually, Luigi is similar to GNU Make where certain **tasks exist which may have dependencies on other tas**ks

- Luigi takes care of a lot of the **workflow management**

- We have adapted Luigi to use the Message Passing Interface (MPI) for parallel processes execution on the HPC.

- Use of Luigi enables execution of **embarrassingly parallel tasks** associated with processing continent-wide processes across the 800+ AGDC tiles.

Contributors to change in an output dataset



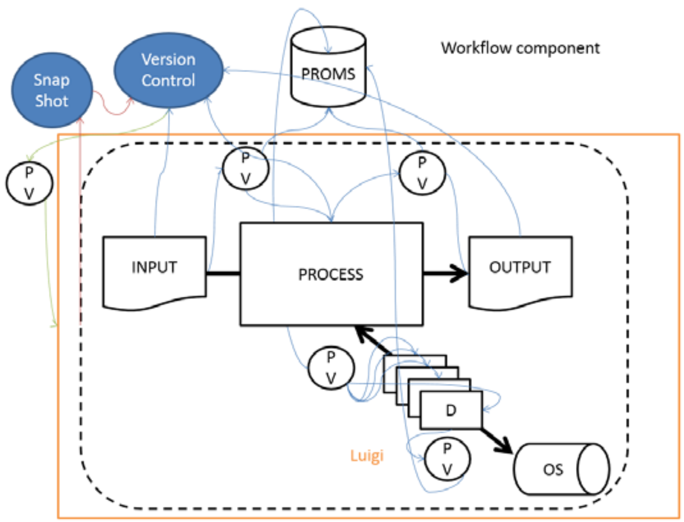
- **Ancillary data version update**
  - Change in geometric base (i.e. image chips used in rectification)
  - Correction parameter update
  - Improved Terrain Model
- **Database schema update**
- **Database content update**
- **Software update**
- **Software library update**
- **Configuration change (command line configuration)**
- **Runtime environment**
  - Operating System
  - Processor architecture
  - Network distribution, if using parallelization
- **Build configuration**
  - Compilation options
- **Change in data model or output format**

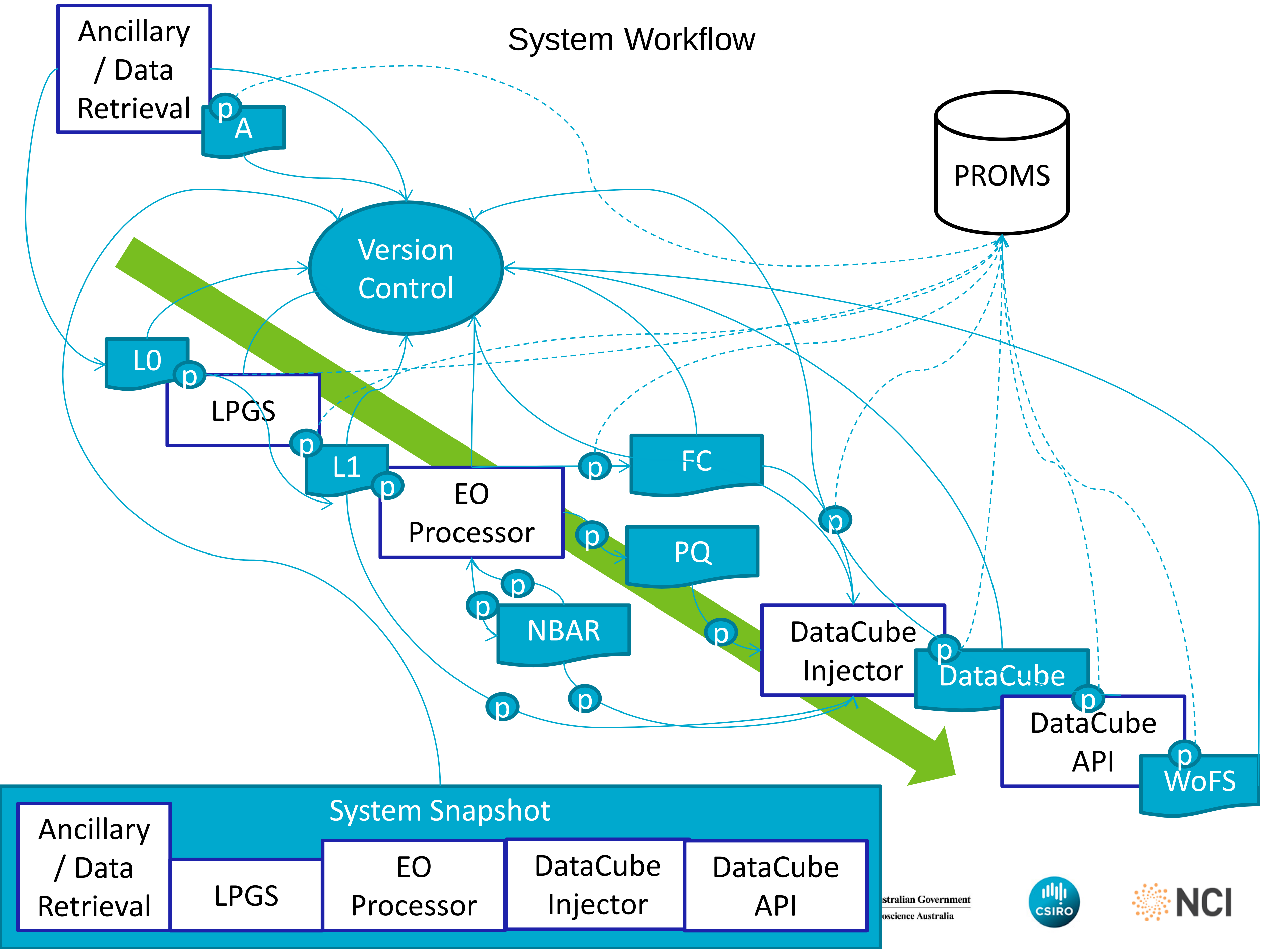
Australian Government Geoscience Australia
CSIRO
NCI

Managing the data collection

Towards repeatable and transparent processes:

- **System Snapshot** as part of Production Rollout
- **Version Control** – software and data
- Automated retrieval of ancillaries and update
- **Provenance reporting** based on version-controlled inputs and outputs
- **Provenance analysis** (relating entities)
- Workflows for automation
- **Patch and reprocess** – task dependencies in workflow used to repair collections



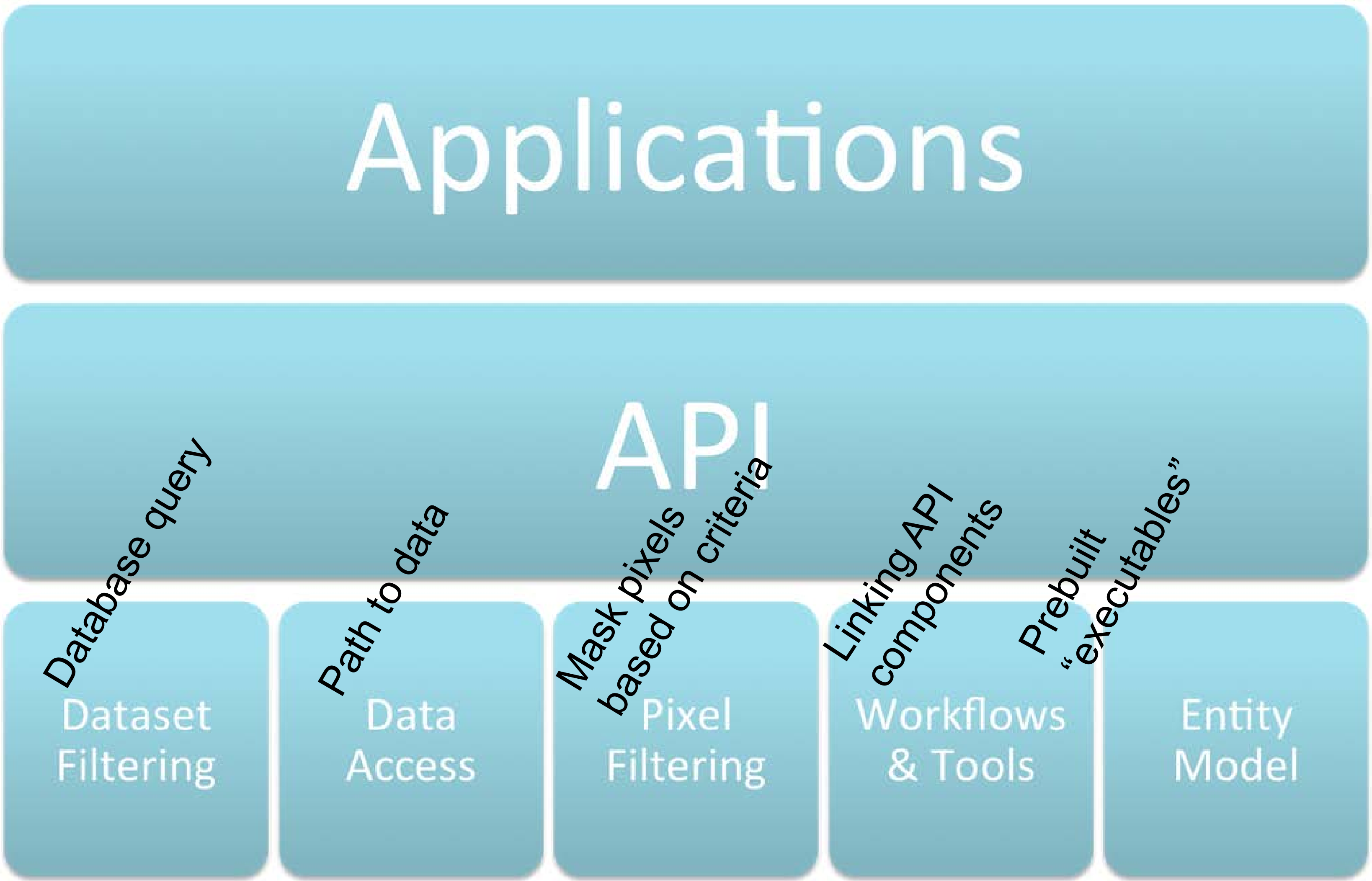

System Workflow
Ancillary / Data Retrieval
p
A
PROMS
Version Control
L0
p
LPGS
p
L1
p
EO Processor
p
FC
p
PQ
p
NBAR
p
p
p
DataCube Injector
p
DataCube
p
DataCube API
p
WoFS
p
p
System Snapshot
Ancillary / Data Retrieval
LPGS
EO Processor
DataCube Injector
DataCube API
NCI

Tasks Underway (or Completed) with Current AGDC

- Code now **open-sourced** on GitHub (https://github.com/GeoscienceAustralia/agdc)
- New release in May(https://github.com/GeoscienceAustralia/agdc/releases)
- Ingesting new data collections using generic ingestion framework (e.g. MODIS).
- Hardening remaining prototype code and optimising prototype DB schema.
- **Developing new APIs** to support specific use case patterns.
- Developing **generic workflow tools** to manage parallel processing (Luigi)
- Delivering basic WMS, WCS, WPS & WCPS **web services**
- Providing simple tools for **cross-sensor interoperability** (e.g. spectral matching/adjustment)

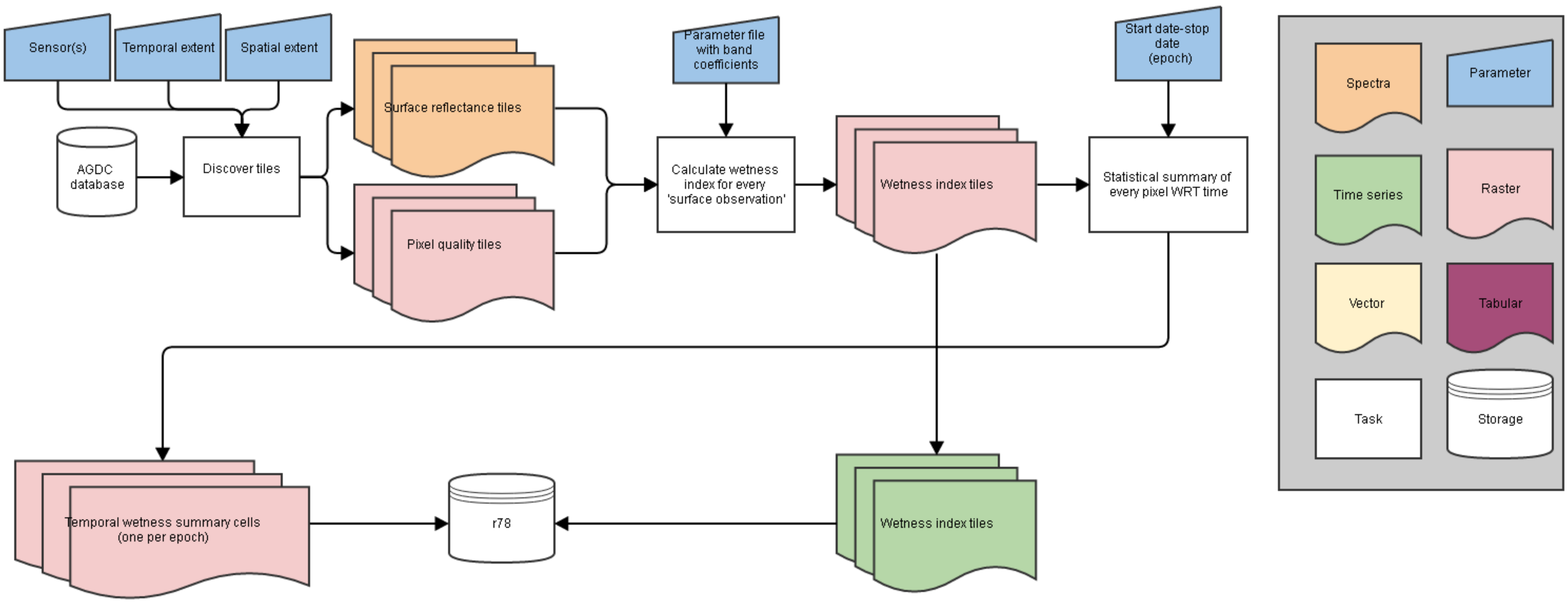
API overview

http://geoscienceaustralia.github.io/agdc/api/0.1.0/html/index.html



Australian Government Geoscience Australia   CSIRO   NCI

Standard Workflow Patterns

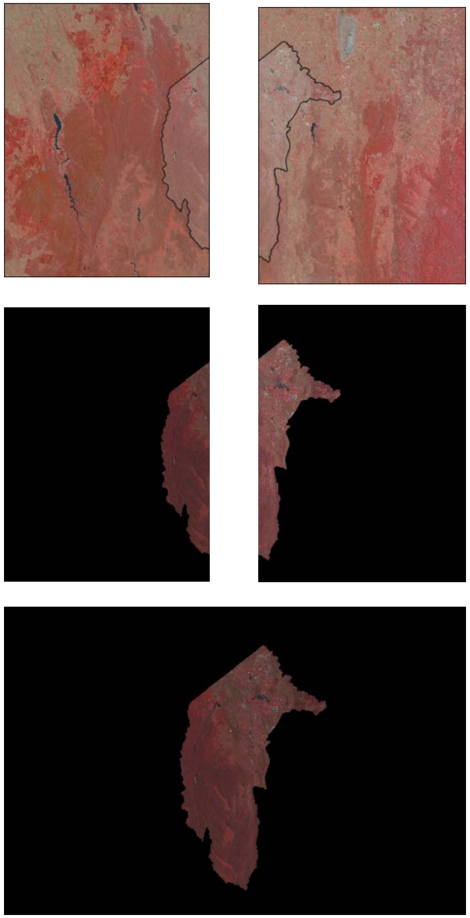


1. Use cases analysed
2. APIs designed
3. Generic, HPC-friendly workflow engines implemented

API command line tools

There are a set of packaged executables for Non-Python "People":

- Retrieve pixel time series
- Retrieve dataset
- Retrieve dataset time series
- Retrieve dataset stack
- Retrieve time series within an AOI
- Summarise dataset time series



Example API execution – pixel drill

```
$ retrieve_pixel_time_series.py -h
usage:                [-h]
                      [--quiet | --verbose]
                      --lat LATITUDE --lon LONGITUDE
                      [--acq-min ACQ_MIN] [--acq-max ACQ_MAX]
                      [--process-min PROCESS_MIN] [--process-max PROCESS_MAX]
                      [--ingest-min INGEST_MIN] [--ingest-max INGEST_MAX]
                      [--satellite {LS5,LS7,LS8} [{LS5,LS7,LS8} ...]]
                      [--apply-pqa]
                      [--pqa-mask
{PQ_MASK_CLEAR,PQ_MASK_SATURATION,PQ_MASK_CONTIGUITY,PQ_MASK_LAND,PQ_MASK_CL
OUD,...} [...]]
                      [--hide-no-data]
                      --dataset-type {ARG25,PQ25,FC25,WATER,...}
                      [--delimiter DELIMITER]
                      [--output-directory OUTPUT_DIRECTORY]
                      [--overwrite]
```

Example API execution – pixel drill ARG25

```
$ retrieve_pixel_time_series.py --lon 120.25 --lat -20.25 --acq-min /
 2013-12 --acq-max 2013-12 --satellite LS7 --dataset-type ARG25 --quiet
```

```
SATELLITE,ACQUISITION DATE,BLUE,GREEN,RED,NEAR_INFRARED, /
SHORT_WAVE_INFRARED_1,SHORT_WAVE_INFRARED_2
LS7,2013-12-01 01:58:47.045319,-999,-999,-999,-999,-999,-999
LS7,2013-12-10 01:53:02.625103,-999,-999,-999,-999,-999,-999
LS7,2013-12-17 01:58:47.468905,388,824,1605,2632,3326,2626
LS7,2013-12-26 01:53:05.686238,-999,-999,-999,-999,-999,-999
```

Example API execution – pixel drill WOfS

```
$ retrieve_pixel_time_series.py --lon 120.25 --lat -20.25 --acq-min /
2013-12 --acq-max 2013-12 --satellite LS7 --dataset-type WATER --quiet
```

```
SATELLITE,ACQUISITION DATE,WATER
LS7,2013-12-01 01:58:23,Saturation/Contiguity,2
LS7,2013-12-10 01:52:38,Saturation/Contiguity,2
LS7,2013-12-17 01:58:23,Dry,0
LS7,2013-12-26 01:52:41,Saturation/Contiguity,2
```

Demonstration available

47GB VM – osgeo-live8.0 Ubuntu Linux

Latest AGDC release

2 months of TM and ETM+ data for a 2x2 tile subset

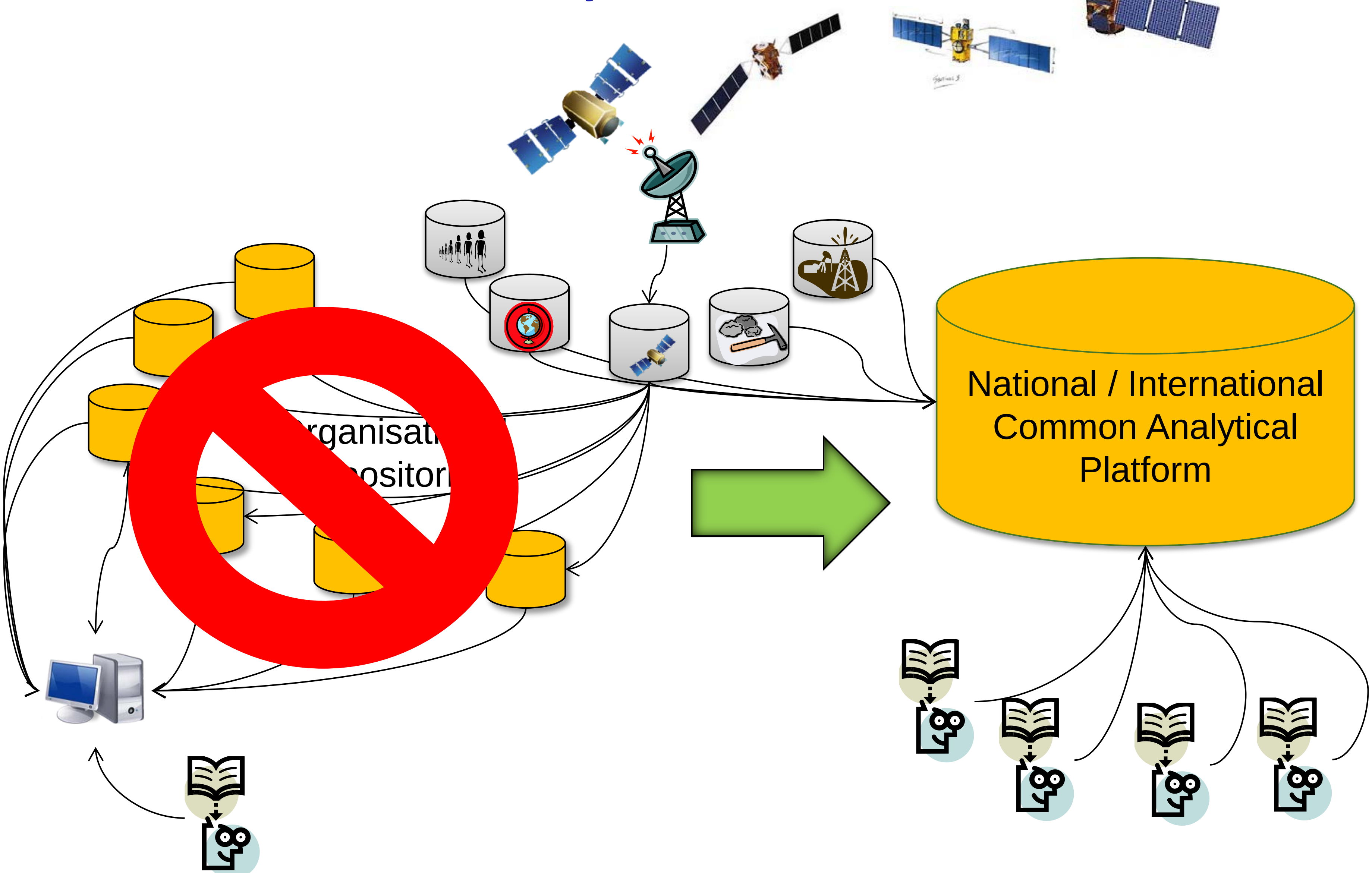
AGDC open source project

https://github.com/GeoscienceAustralia/agdc

Examples

1. Create a tile listing based on input tile and criteria
2. Filter pixels based on quality
3. Create an RGB image from tile contents for a date range
4. Iteratively create multiple indices for a Data Cube tile
5. Submit bulk processing over a selected area

Towards a Common Analytical Platform



What does it cost to make a Data Cube?

Unlocking the Landsat Archive project:	~AUD **$3.5M** over 3 years
Second year of Data Cube :	~AUD **$1.5M**
Current year of Data Cube :	~AUD **$2M**
WOfS first year (prototype application):	~AUD **$1M**
NCI membership	~AUD **$0.5M** for 3 years?

▪http://currencyguide.eu/

Online storage rate of $500/TB/YR currently covered by RDSI funding

Australian Government
Geoscience Australia
CSIRO
NCI

Questions!

Simon.Oliver@ga.gov.au