

IN43D-07

AGU Fall Meeting 2016



NCI

AUSTRALIA

Implementing a Data Quality Strategy to simplify access to data

Kelsey Druken, Claire Trenham, Ben Evans, Clare Richards, Jingbo Wang, & Lesley Wyborn

National Computational Infrastructure, Canberra, Australia

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative



Australian Government
Bureau of Meteorology



Australian Government
Geoscience Australia



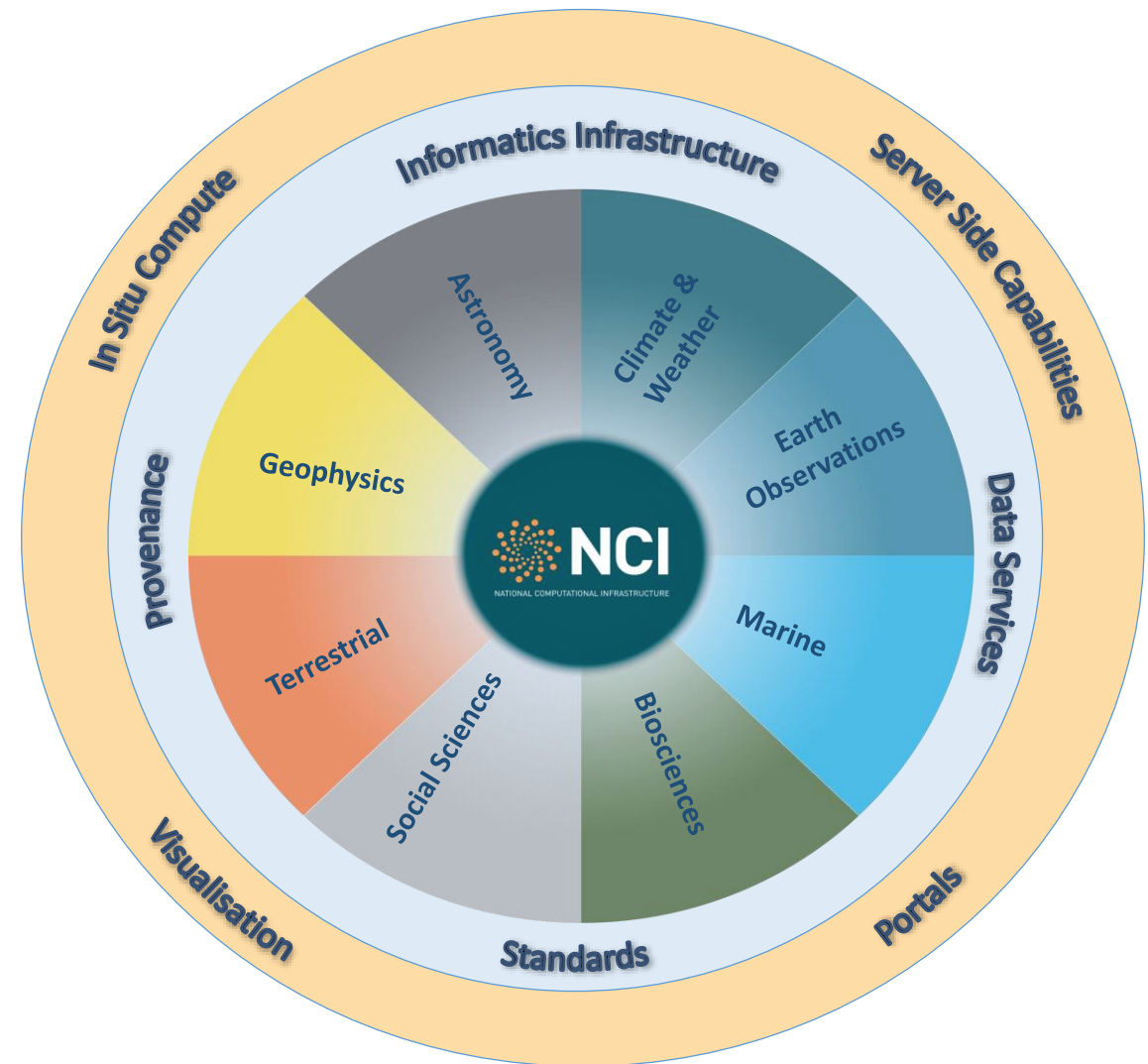
Australian Government
Australian Research Council



Australian
National
University

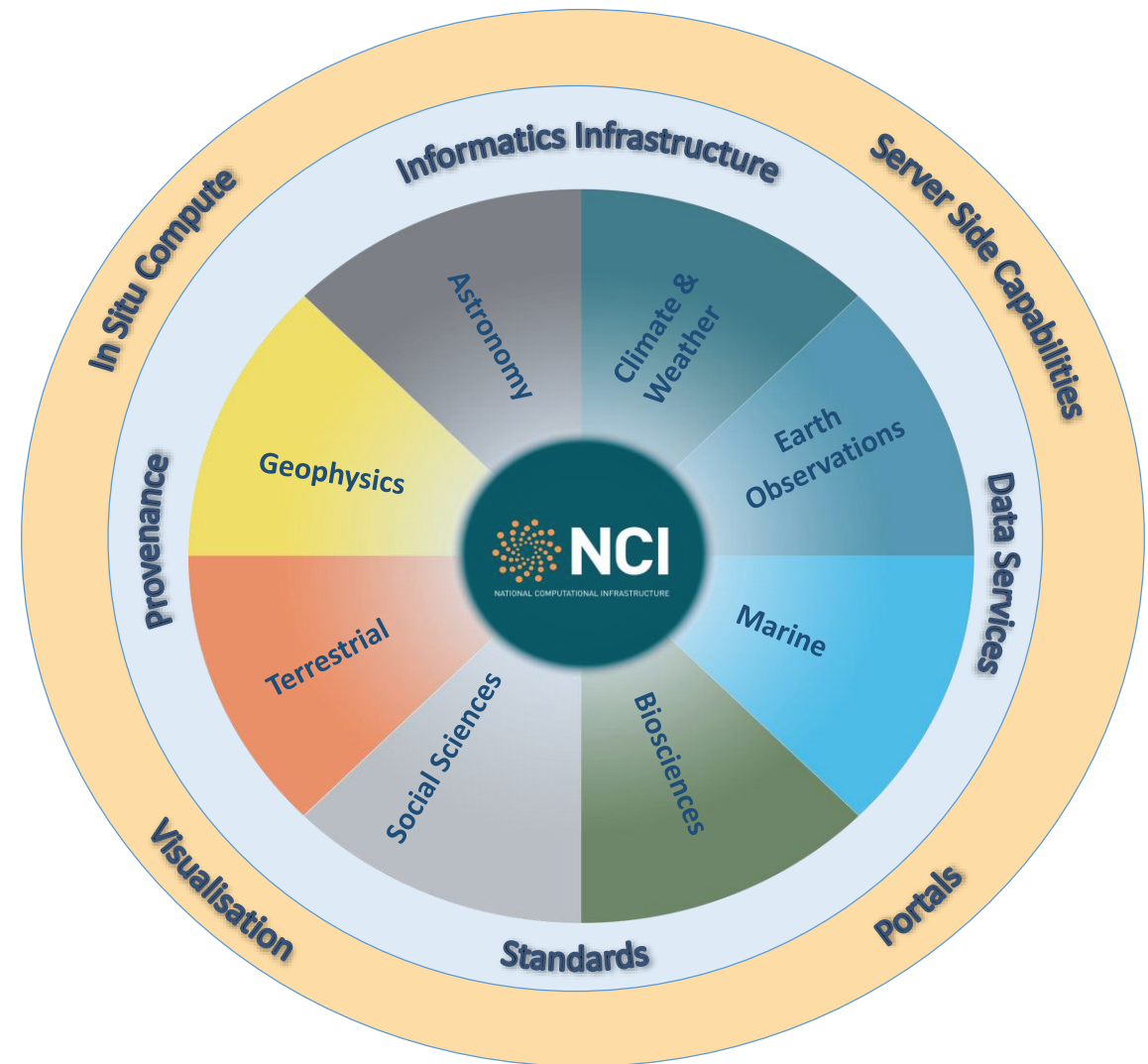
nci.org.au
 @NCInews

- NCI hosts one of Australia's largest repositories (10+ PBytes) of research data collections
- Spanning data collections from climate, coasts, oceans and geophysics through to astronomy, bioinformatics and the social sciences



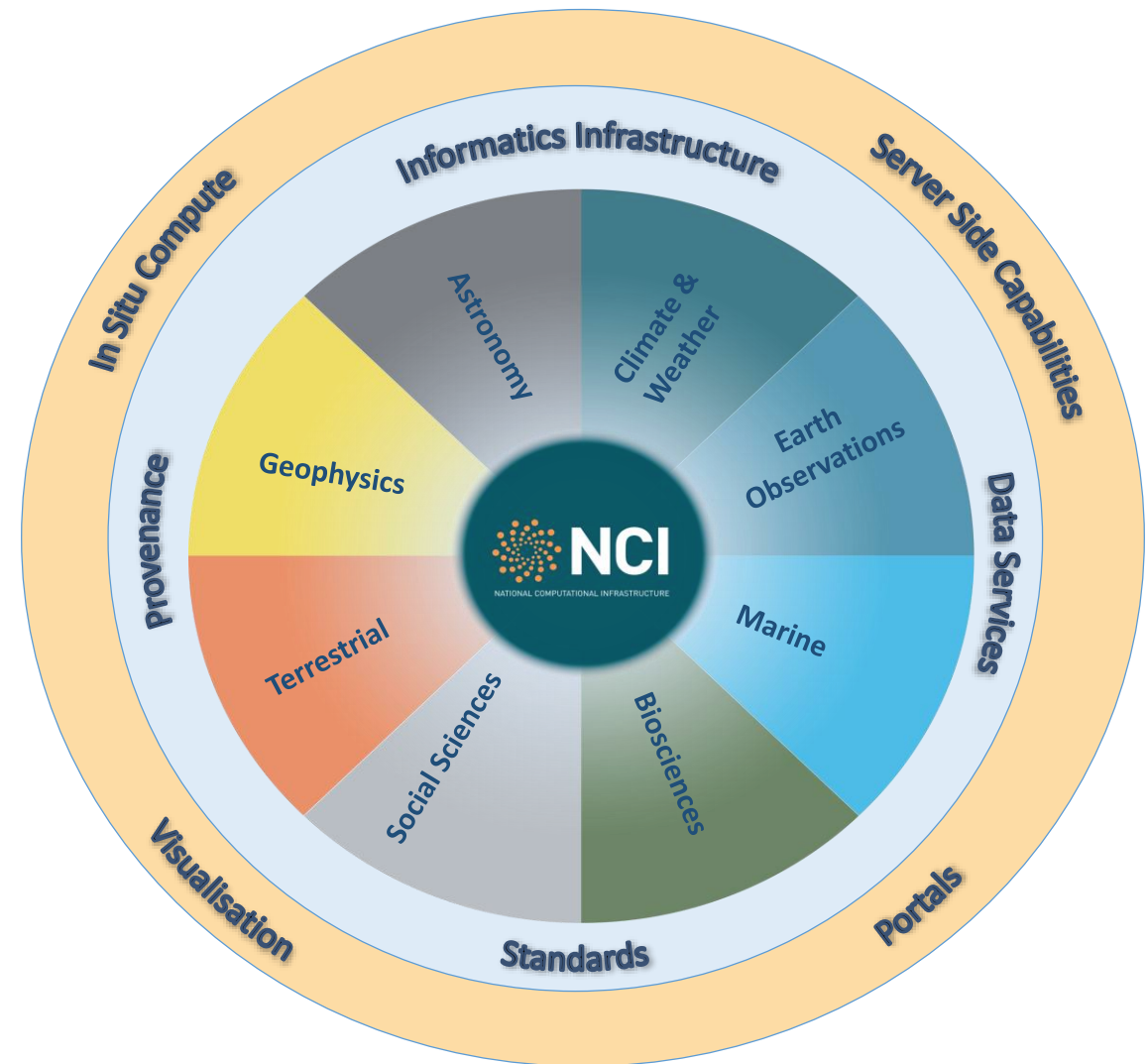
Key to maximizing benefit of NCI's collections and computational capabilities:

→ Ensuring seamless interoperable access to these datasets

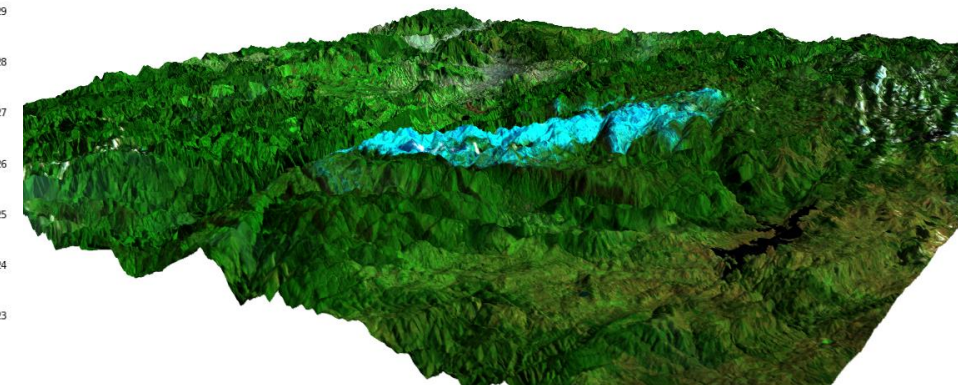
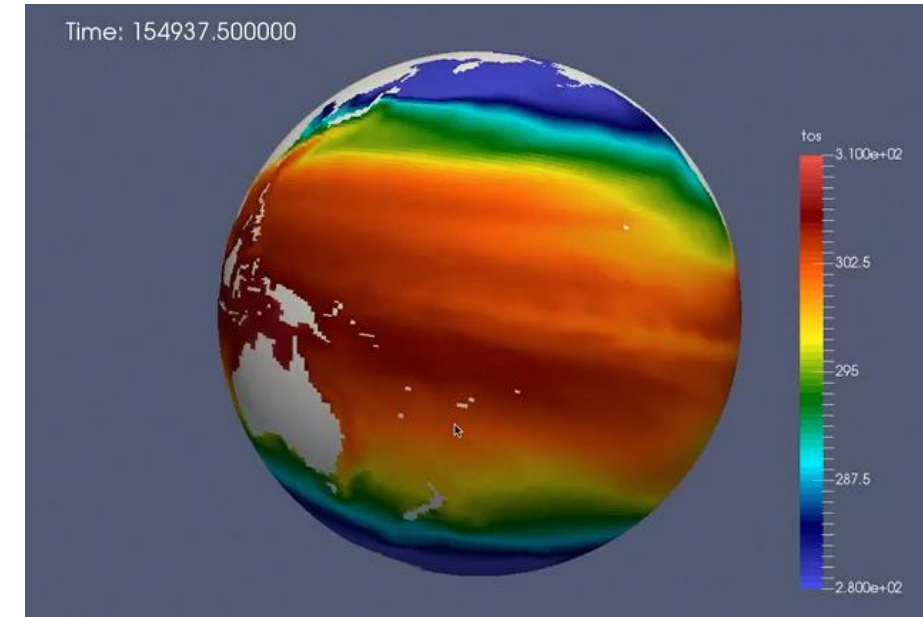
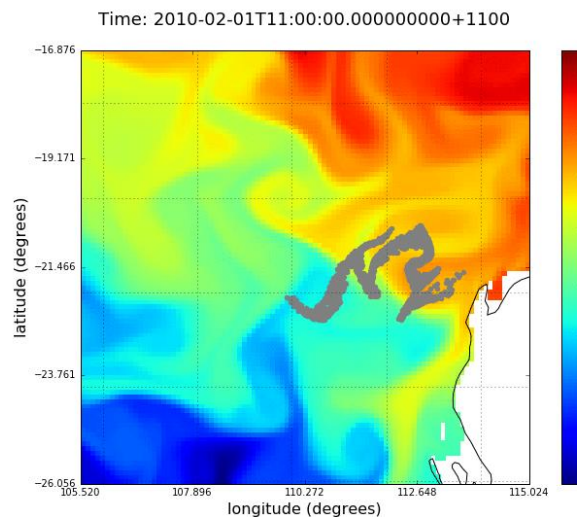
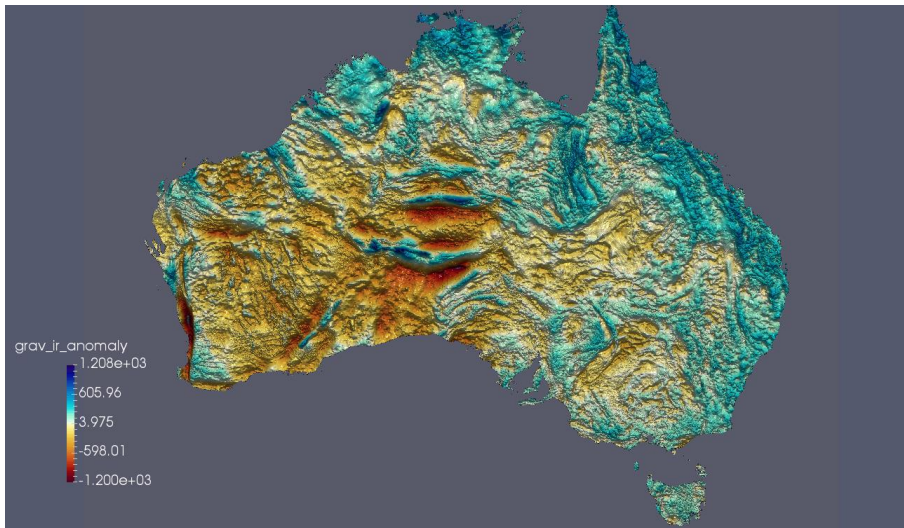


Key to maximizing benefit of NCI's collections and computational capabilities:

→ Ensuring seamless interoperable access to these datasets

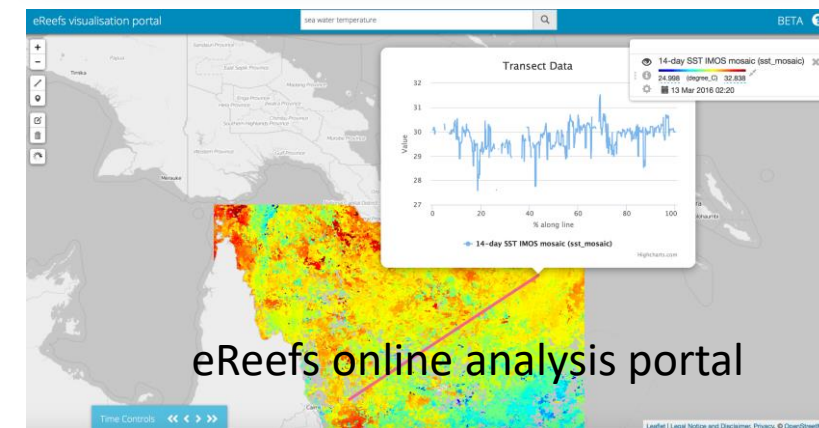
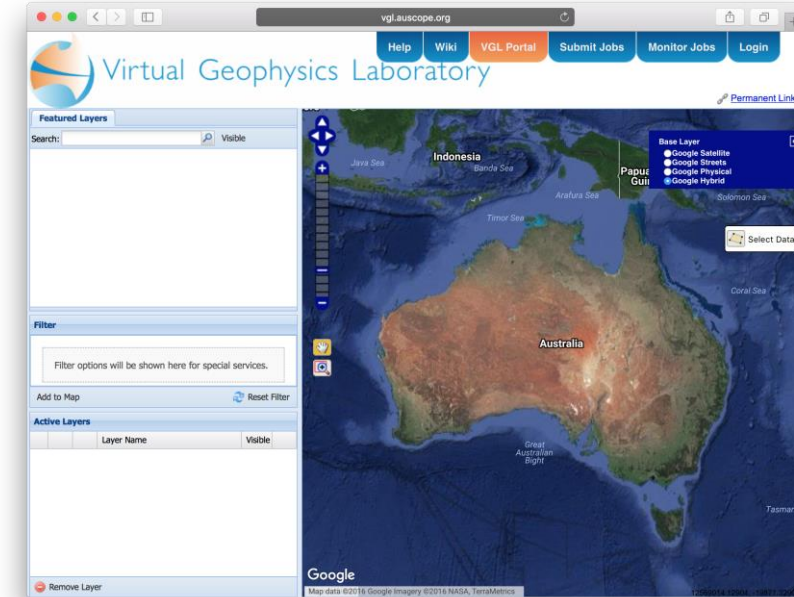
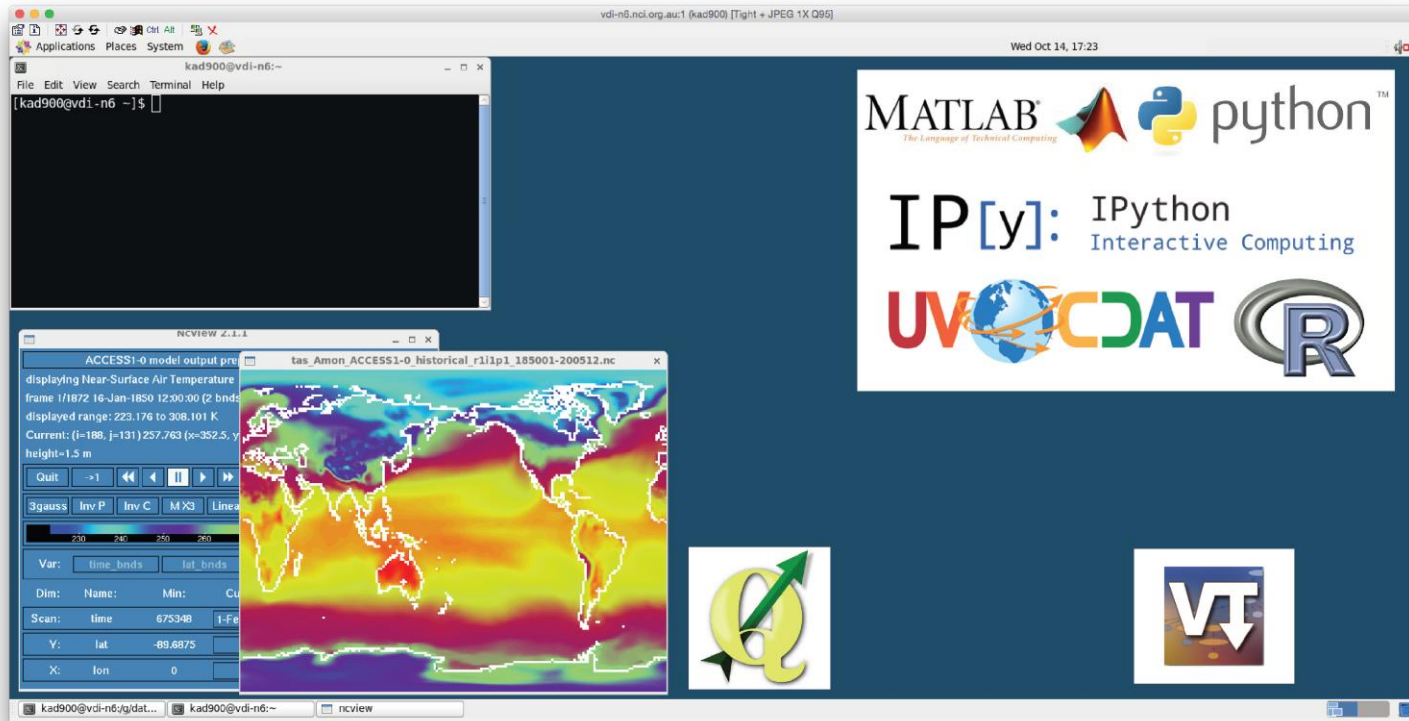


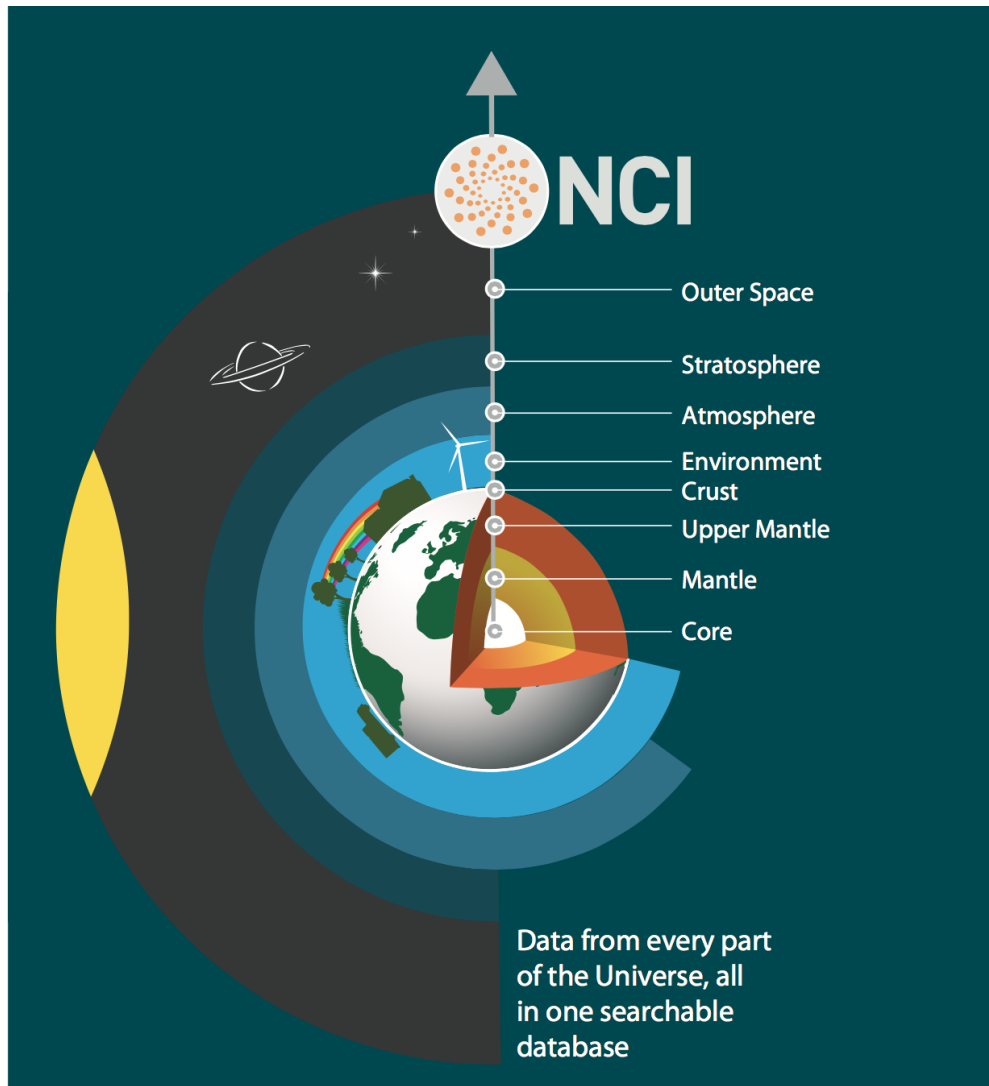
- Combining data
- Visualising
- How can we enable this type of easy access and use?



Collections are being accessed and utilised from a broad range of options

- Direct access on filesystem
- Web and data services
- Data portals
- Virtual labs (e.g., virtual desktops)





- Application of community-agreed data standards to the broad set of Earth systems and environmental data that are being used
- Within these disciplines, data span a wide range of:
 - Gridded
 - Non-gridded (i.e., trajectories/profiles, point data)
 - Coordinate reference projections
 - Resolutions





Shelley Stall
Assistant Director,
Enterprise Data
Management Program

<http://dataservices.agu.org/dmm/>

DMM Capability – 25 Processes to Perform, Manage, Define

1. Data Management Strategy Process Area

1. Data Management Strategy
2. Communications
3. Data Management Function
4. Grant Strategy/Business Case
5. Funding

2. Data Governance Process Area

6. Governance Management
7. Vocabulary/Glossary
8. Metadata Management

3. Data Quality Process Area

9. Data Quality Strategy
10. Data Profiling
11. Data Quality Assessment
12. Data Cleansing and Curation

4. Data Operations Process Area

13. Data Requirements Definition
14. Data Lifecycle Management
15. Contribution / Provider Management

5. Platform and Architecture Process Area

16. Architectural Standards
17. Architectural Approach
18. Data Management Platform
19. Data Integration / Data Linking
20. Data Archiving and Preservation

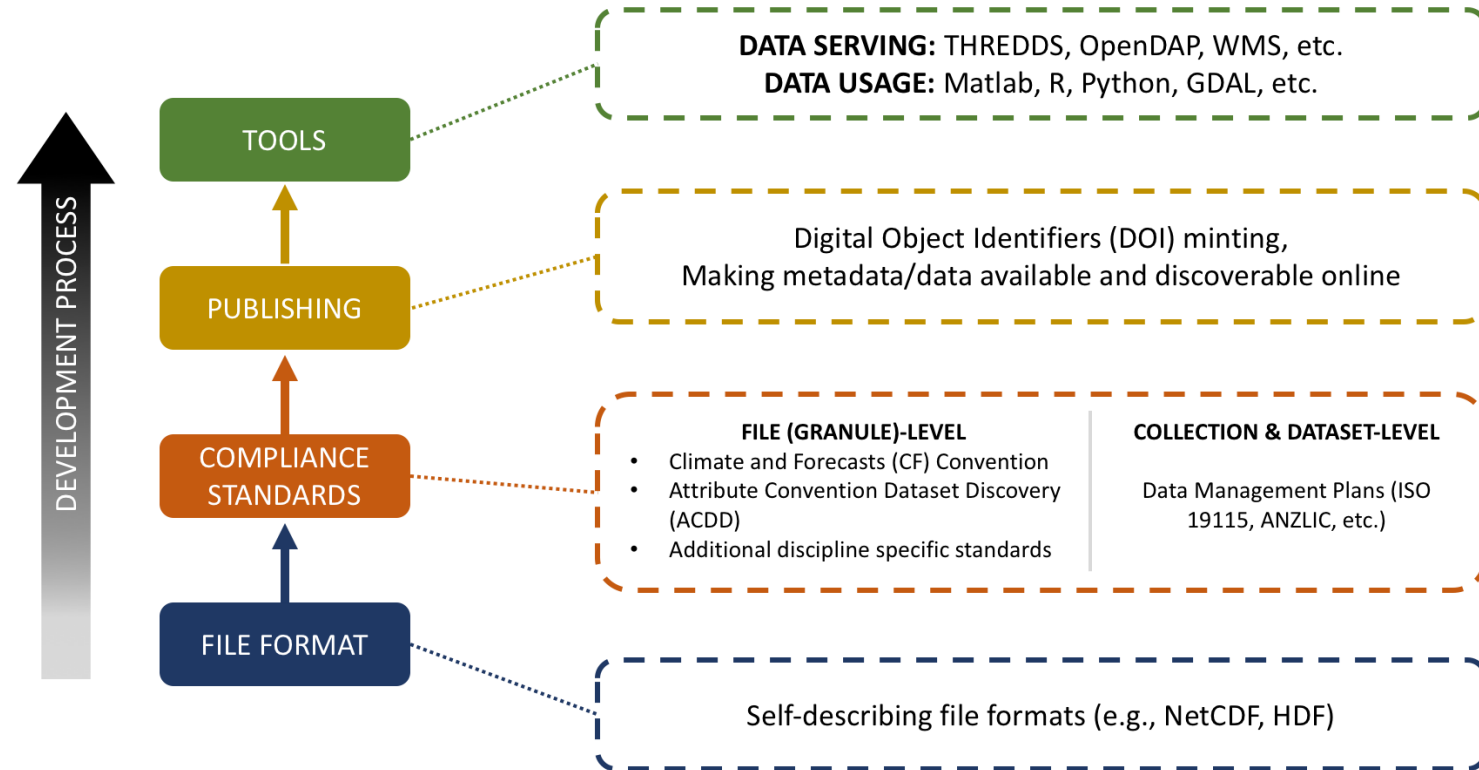
6. Infrastructure Support Practices

21. Measurement and Analysis
22. Process Management
23. Process Quality Assurance
24. Risk Management
25. Configuration Management



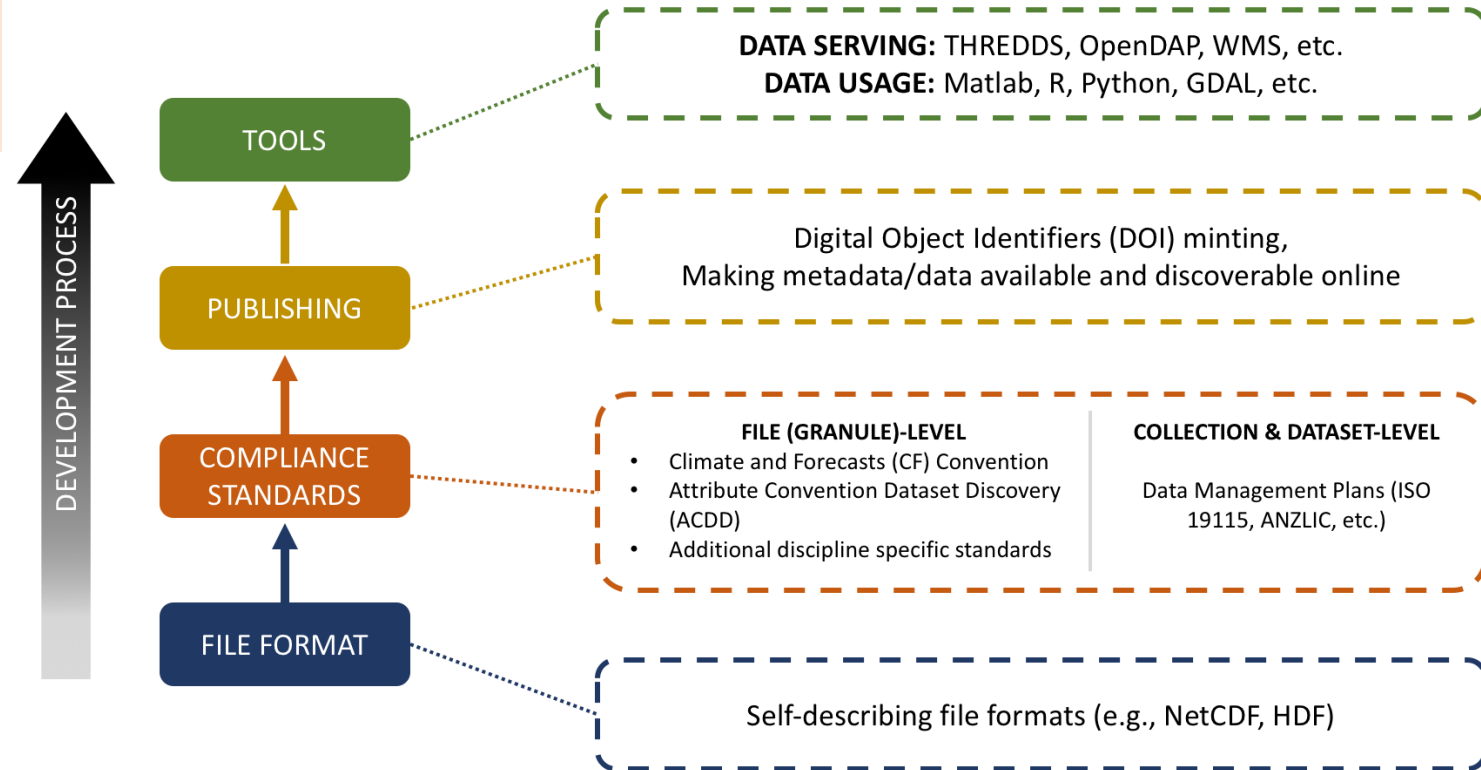
Data Quality Strategy (DQS): What does it involve?

1. Underlying High Performance Data (HPD) file format
2. Close collaboration with data custodians and managers
 - Planning, designing, and assessing the data collections
3. Quality control through compliance with recognised community standards
4. Data assurance through demonstrated functionality across common platforms, tools, and services

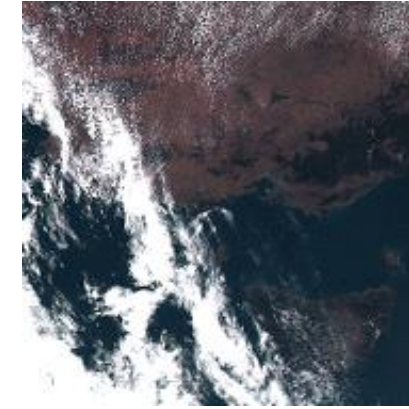
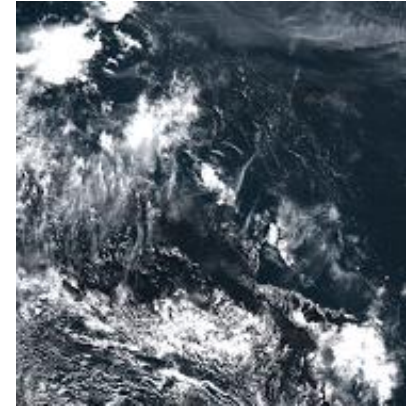
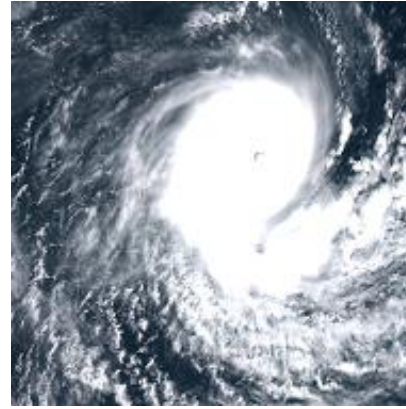


Data Quality Strategy (DQS): What does it involve?

1. Underlying High Performance Data (HPD) file format
2. Close collaboration with data custodians and managers
 - Planning, designing, and assessing the data collections
3. Quality control through compliance with recognised community standards
4. Data assurance through demonstrated functionality across common platforms, tools, and services



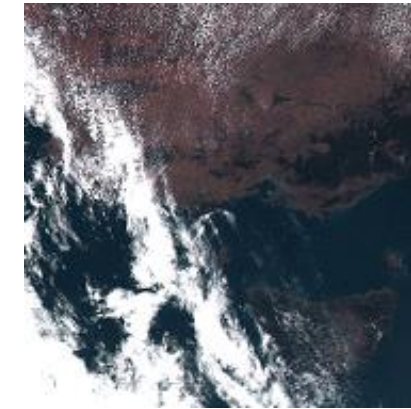
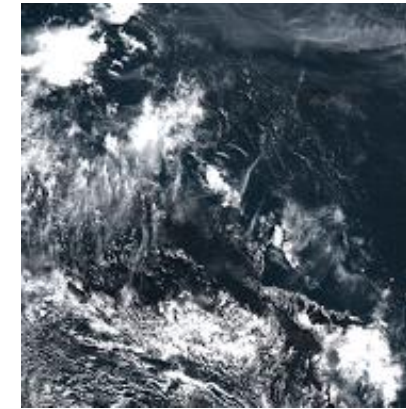
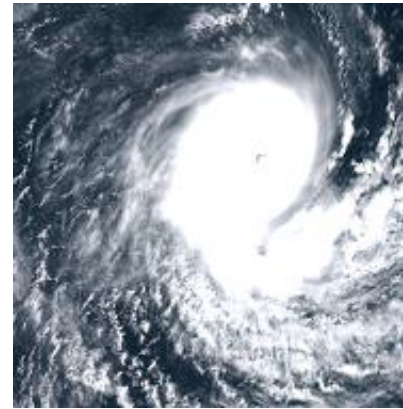
1. Climate/ESS Model Assets and Data Products
2. Earth and Marine Observations and Data Products
3. Geoscience Collections
4. Terrestrial Ecosystems Collections
5. Water Management and Hydrology Collections










Data Collections	Approx. Capacity
CMIP5, CORDEX, ACCESS Models	5 Pbytes
Satellite Earth Obs: LANDSAT, Himawari-8, Sentinel, MODIS, INSAR	2 Pbytes
Digital Elevation, Bathymetry Onshore/Offshore Geophysics	1 Pbytes
Seasonal Climate	700 Tbytes
Bureau of Meteorology Observations	350 Tbytes
Bureau of Meteorology Ocean-Marine	350 Tbytes
Terrestrial Ecosystem	290 Tbytes
Reanalysis products	100 Tbytes



1. Climate/ESS Model Assets and Data Products
2. Earth and Marine Observations and Data Products
3. Geoscience Collections
4. Terrestrial Ecosystems Collections
5. Water Management and Hydrology Collections

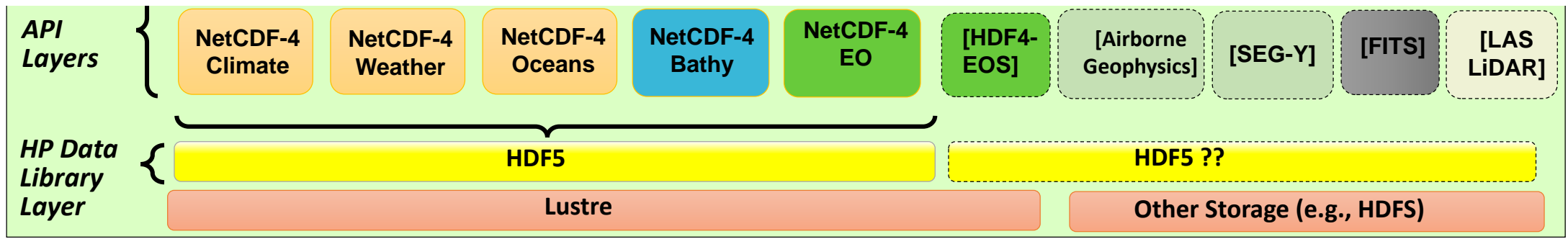


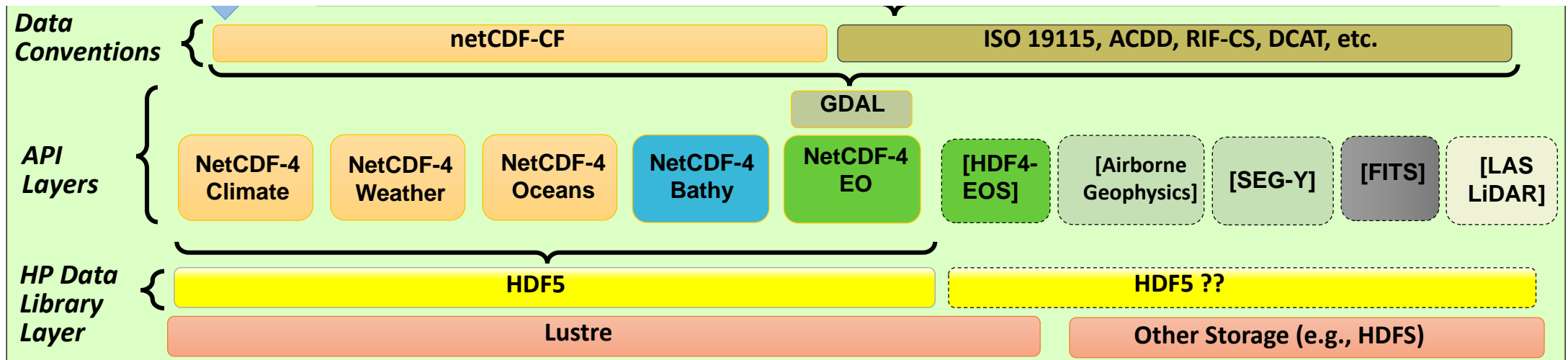
Data Collections	Approx. Capacity
CMIP5, CORDEX, ACCESS Models	5 Pbytes
Satellite Earth C  INSAR	2 Pbytes
Digital Elevator Onshore/Offshc 	1 Pbytes
Seasonal Climat 	700 Tbytes
Bureau of Mete 	350 Tbytes
Bureau of Mete 	350 Tbytes
Terrestrial Ecosy 	290 Tbytes
Reanalysis products 	100 Tbytes

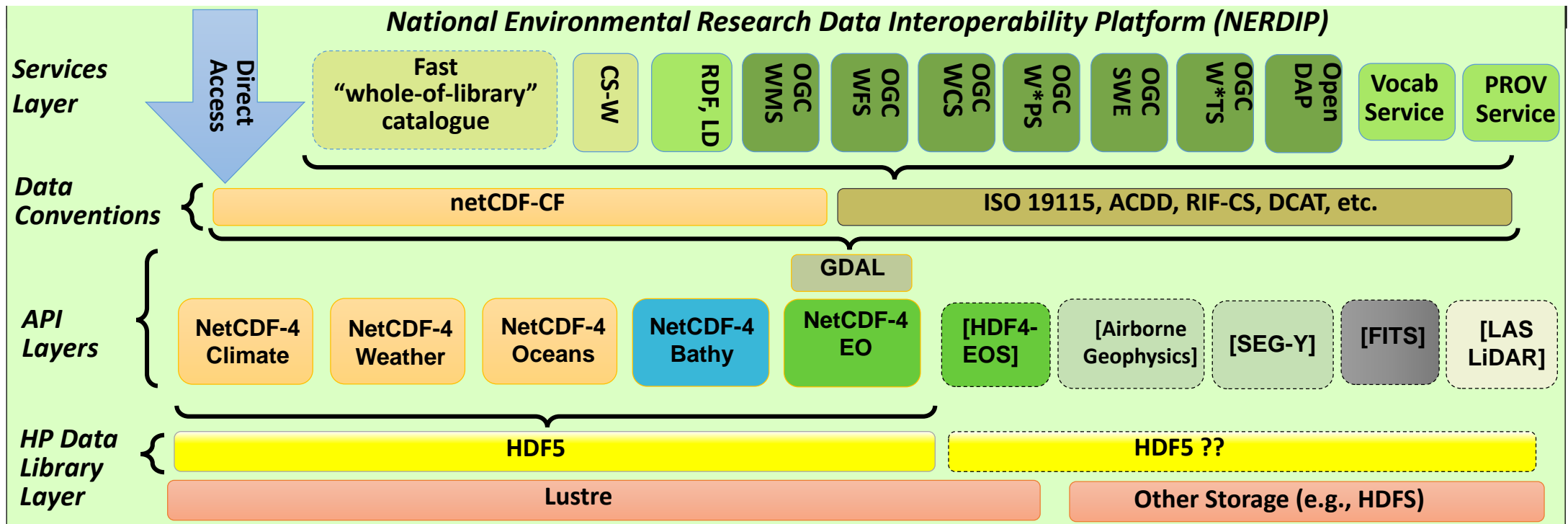
NetCDF
common data format

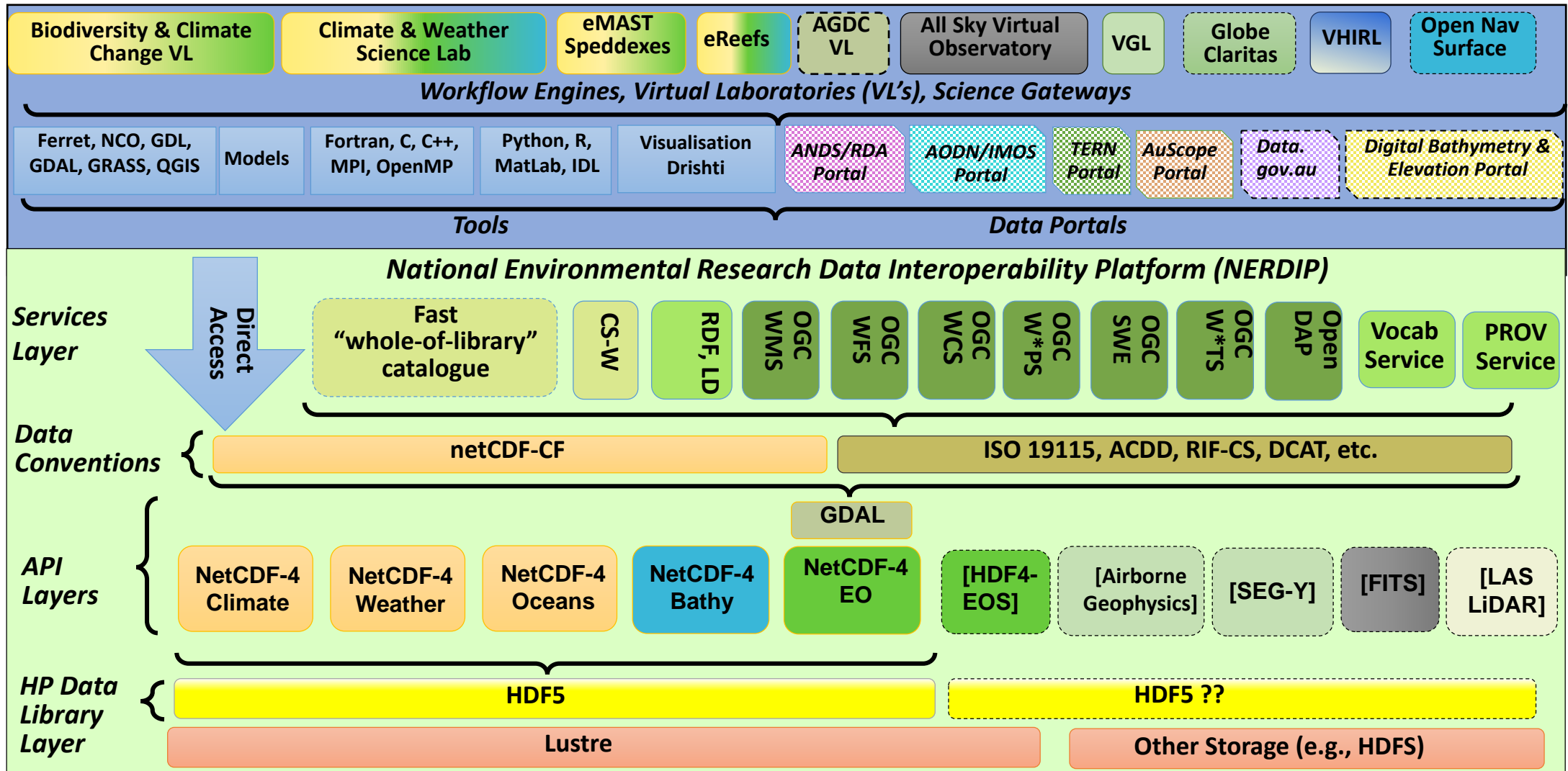






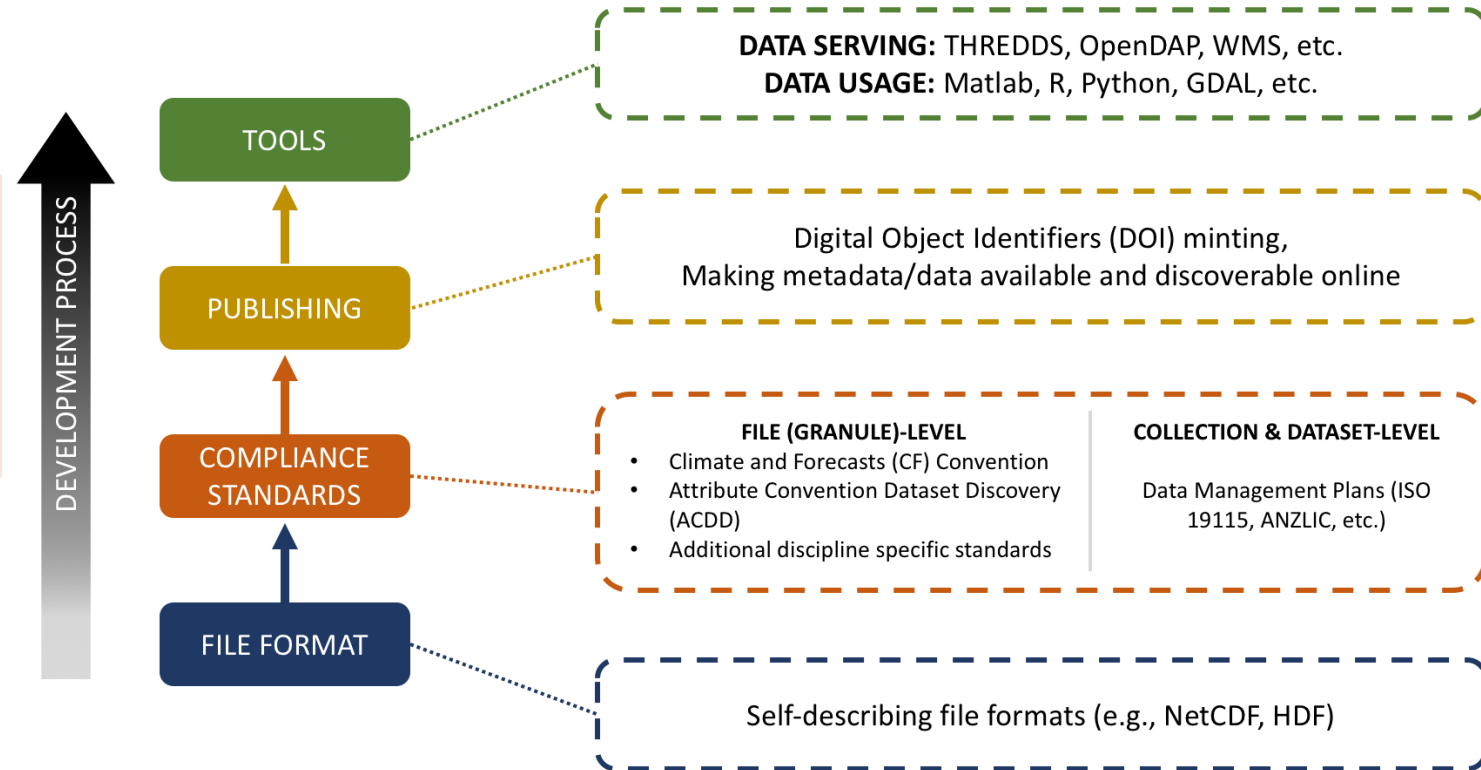






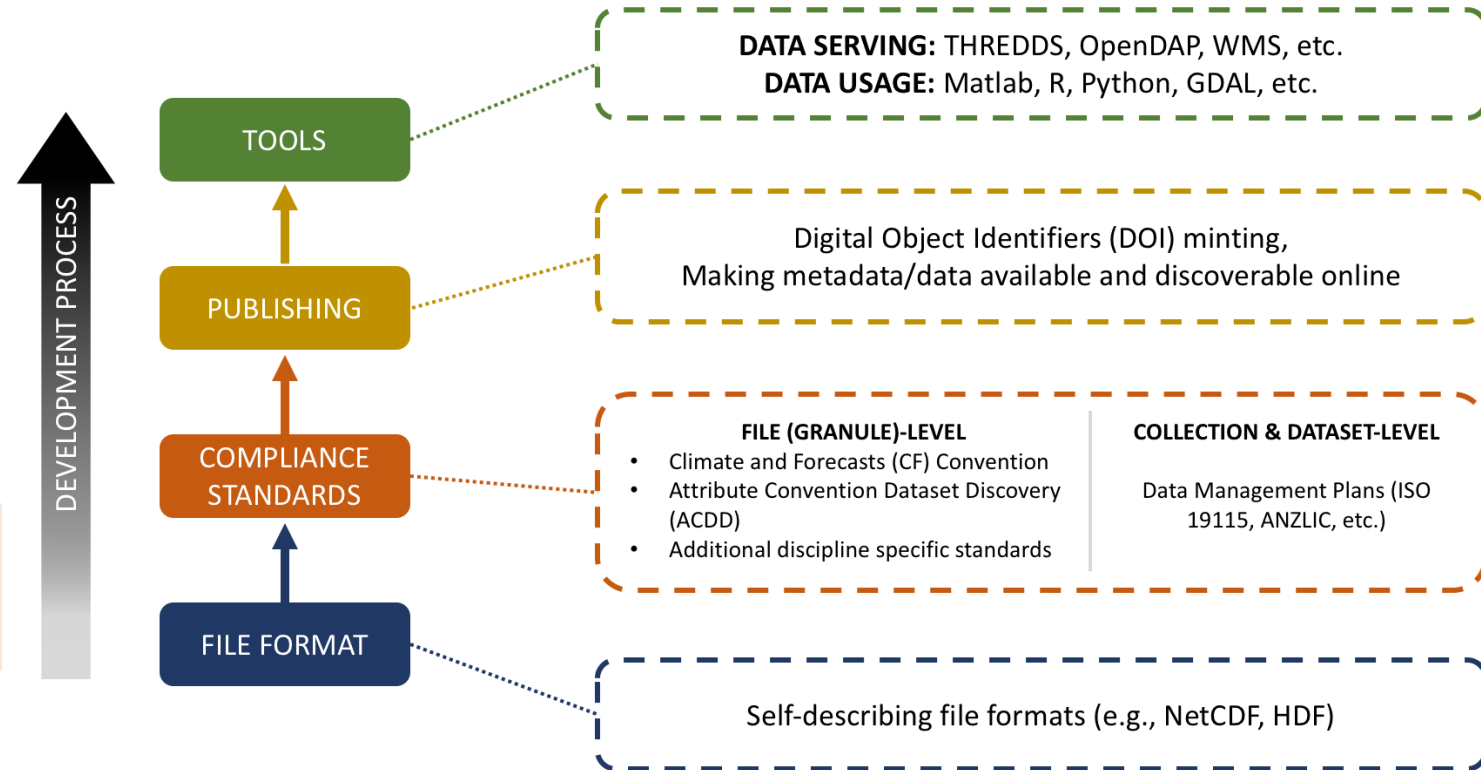
Data Quality Strategy (DQS): What does it involve?

1. Underlying High Performance Data (HPD) file format
2. Close collaboration with data custodians and managers
 - Planning, designing, and assessing the data collections
3. Quality control through compliance with recognised community standards
4. Data assurance through demonstrated functionality across common platforms, tools, and services



Data Quality Strategy (DQS): What does it involve?

1. Underlying High Performance Data (HPD) file format
2. Close collaboration with data custodians and managers
 - Planning, designing, and assessing the data collections
3. Quality control through compliance with recognised community standards
4. Data assurance through demonstrated functionality across common platforms, tools, and services



Collection & dataset-levels

(e.g., parent-child metadata)

ISO-19115, ANZLIC, etc.

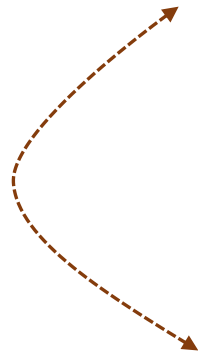
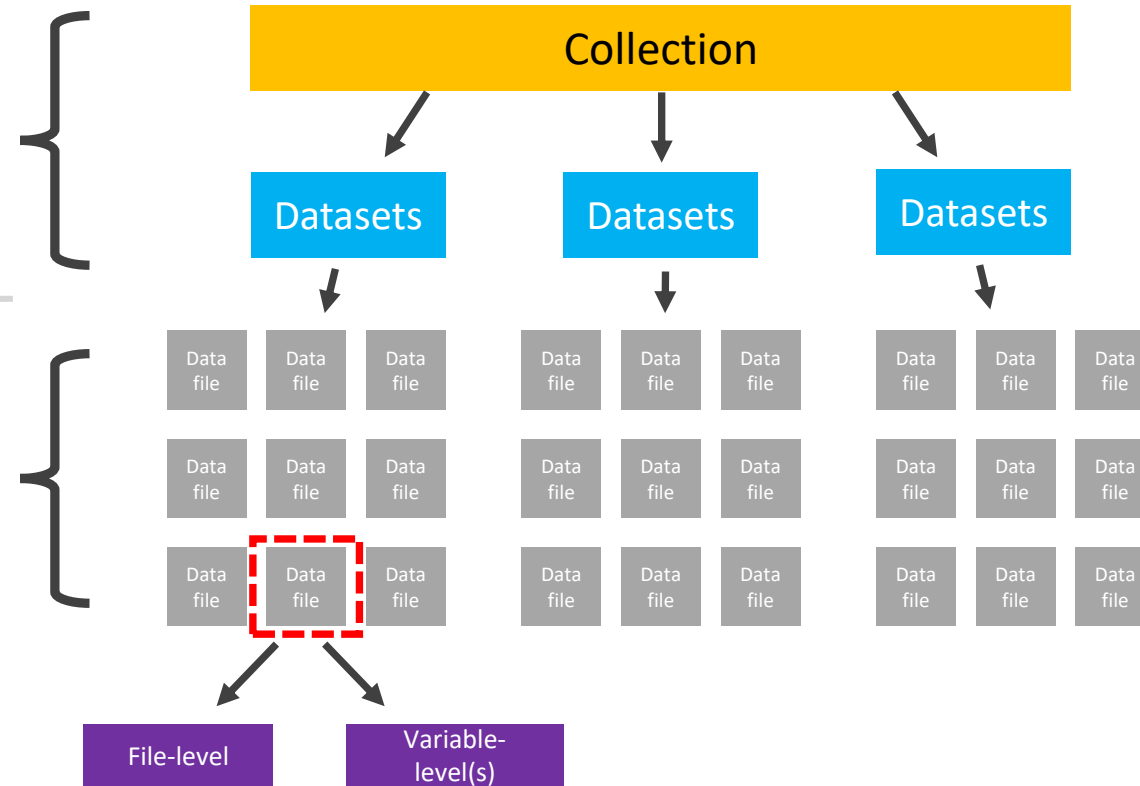
File (granule)-level

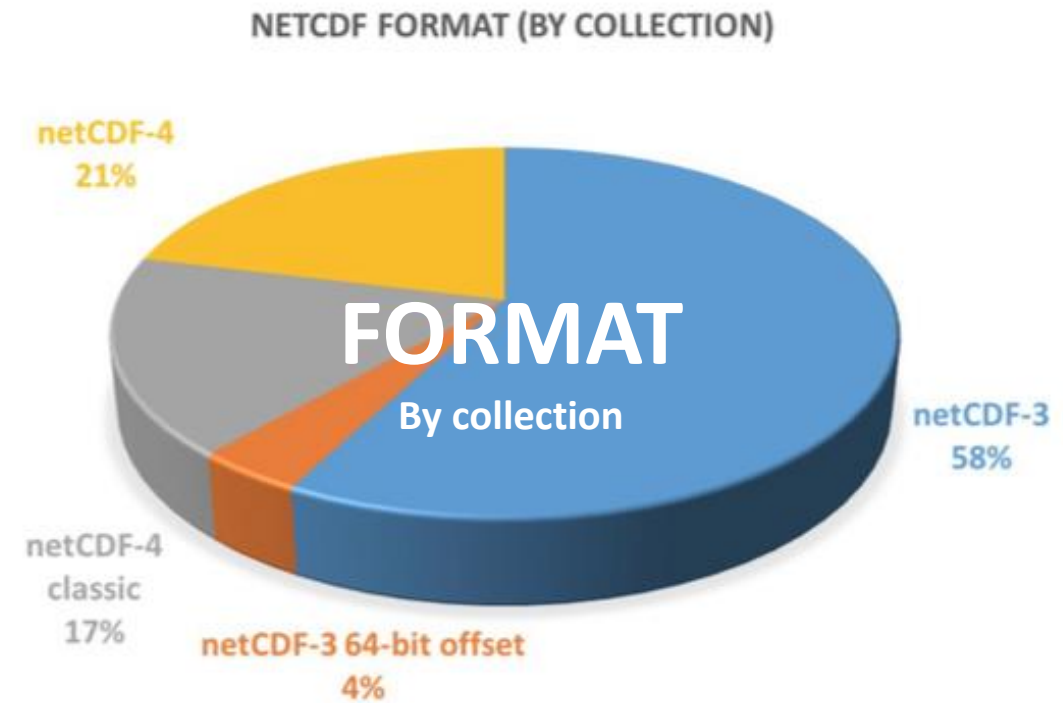
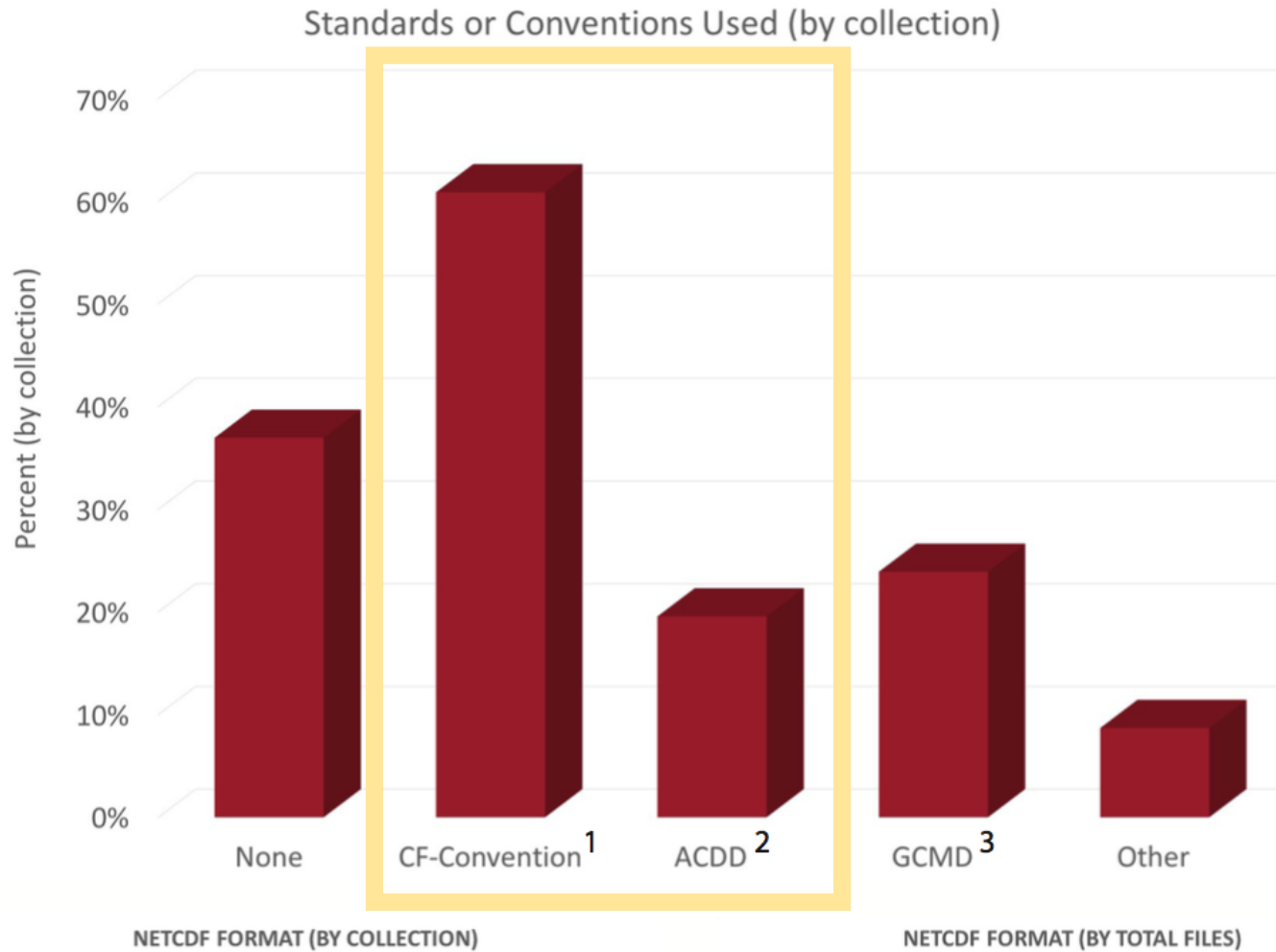
Contains 2 types of metadata:

(1) Variable-level (CF-Convention)

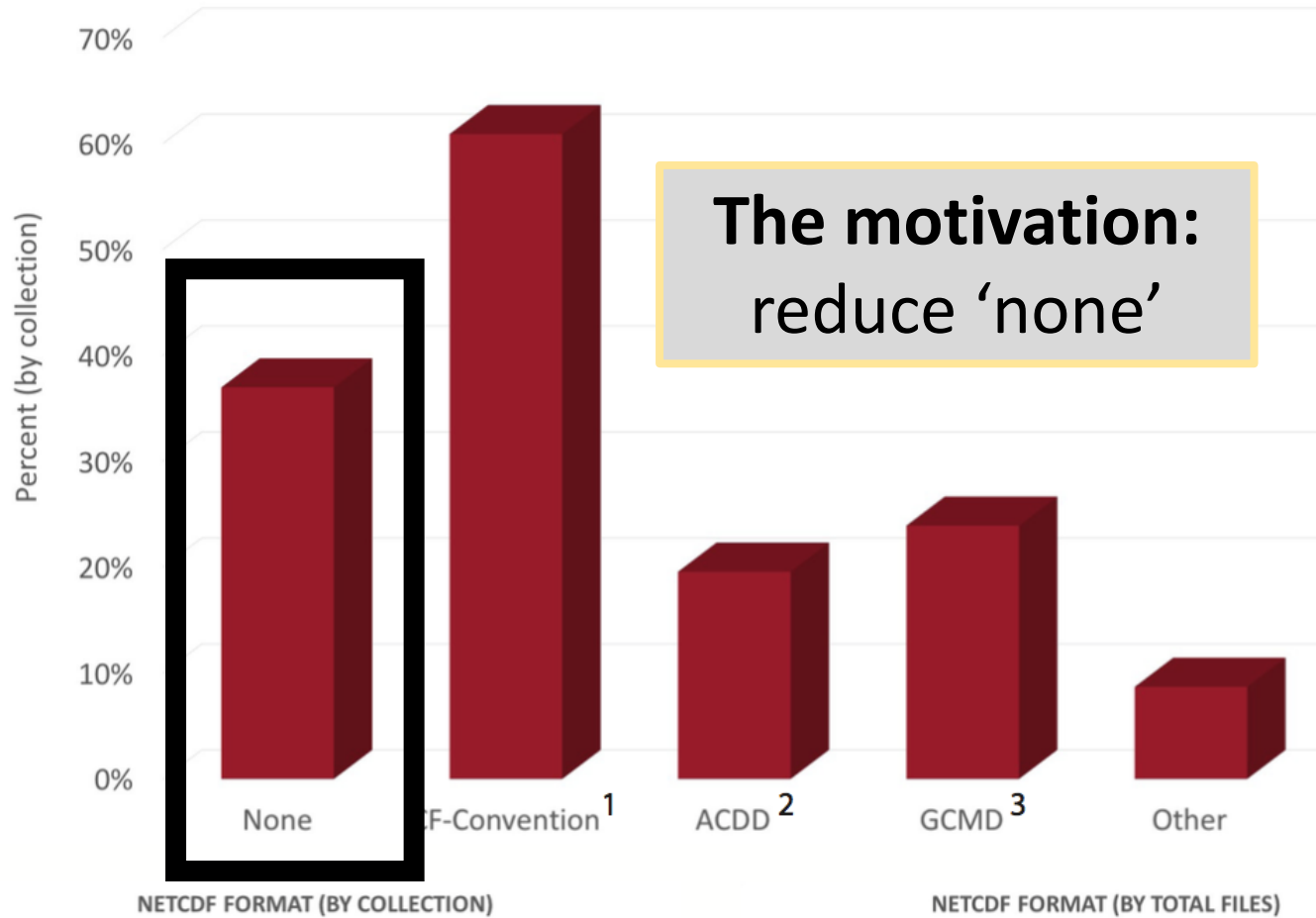
(2) Global file-level (ACDD**)

**Can link to collection/dataset metadata

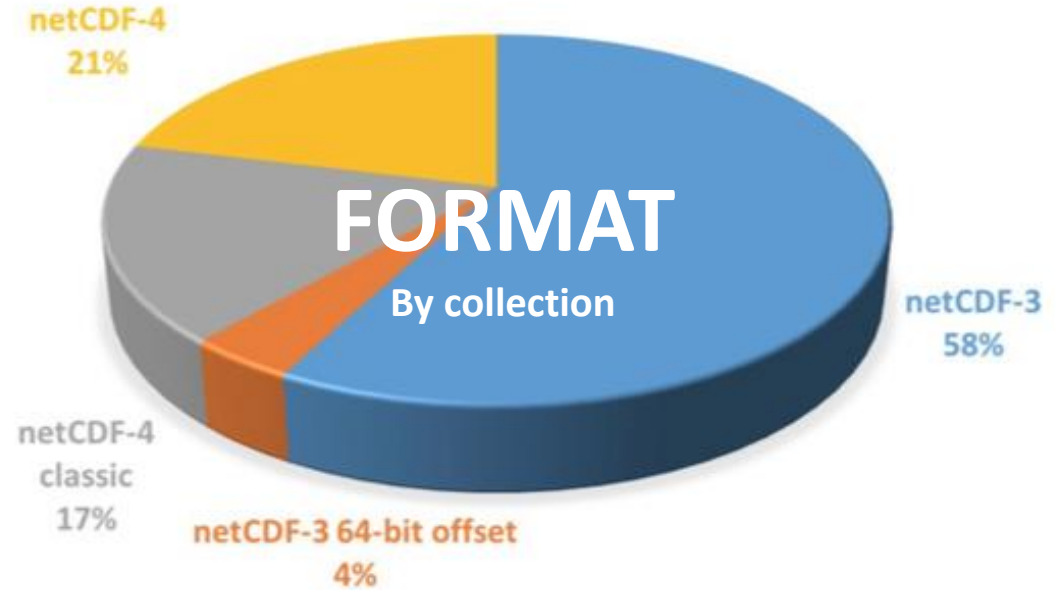




Standards or Conventions Used (by collection)



NETCDF FORMAT (BY COLLECTION)



CF Conventions

- Climate and Forecast Conventions and Metadata: <http://cfconventions.org/>

ACDD

- Attribute Convention for Data Discovery:
[http://wiki.esipfed.org/index.php/Attribute Convention for Data Discovery](http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery)

Together, these two standards define several categories of metadata ensuring:

Usage, discoverability, and understanding of the data contents



- Want to adopt or utilise existing community checkers if possible
- Two main options:
 - **UK Reading (CF-Convention website links to this one)**
 - IOOS (growing fast, designed to be modified and extended)

Our own modifications

- Needed our own wrapper to enable collection-level scans
- Tailor our output and reporting





NCI Quality Control: NetCDF Compliance Report

COLLECTION: [ENTER COLLECTION NAME]

LOCATION: [COLLECTION LOCATION]

Overall comments:

[Brief overall status/report]

Notes/Reminder(s):

The QC report and feedback does not address file performance. Performance tests will be completed separately and in some cases may require additional changes to the CF metadata.

For optimal display of Web Map Services, please consider providing NCI Data Services with an appropriate [min/max] colour scale range for geospatial gridded data content.

Compliance Scoring (report attached):

Total Files Checked	
Total Files Skipped	

	CF* v1.6	ACDD** v1.3	Completeness***
Required elements			--
Additional Metadata	--	--	
File format(s) used	--	--	
Convention(s) used	--	--	

* Climate and Forecast Metadata Convention

** Attribute Convention for Data Discovery

*** Indicators of consistency across the collection or subcollection

High-priority suggestions (for CF and ACDD compliance):

[LIST]

Medium-priority suggestions:

[LIST]

Low-priority suggestions:

[LIST]

Compliance checker

Summarised version on the compliance status.

The break down... compliance scores and also measure of consistency across the collection

Providing attack plan for improvements:
Make it easy for data managers to efficiently address and meet baseline compliance

oken@anu.edu.au

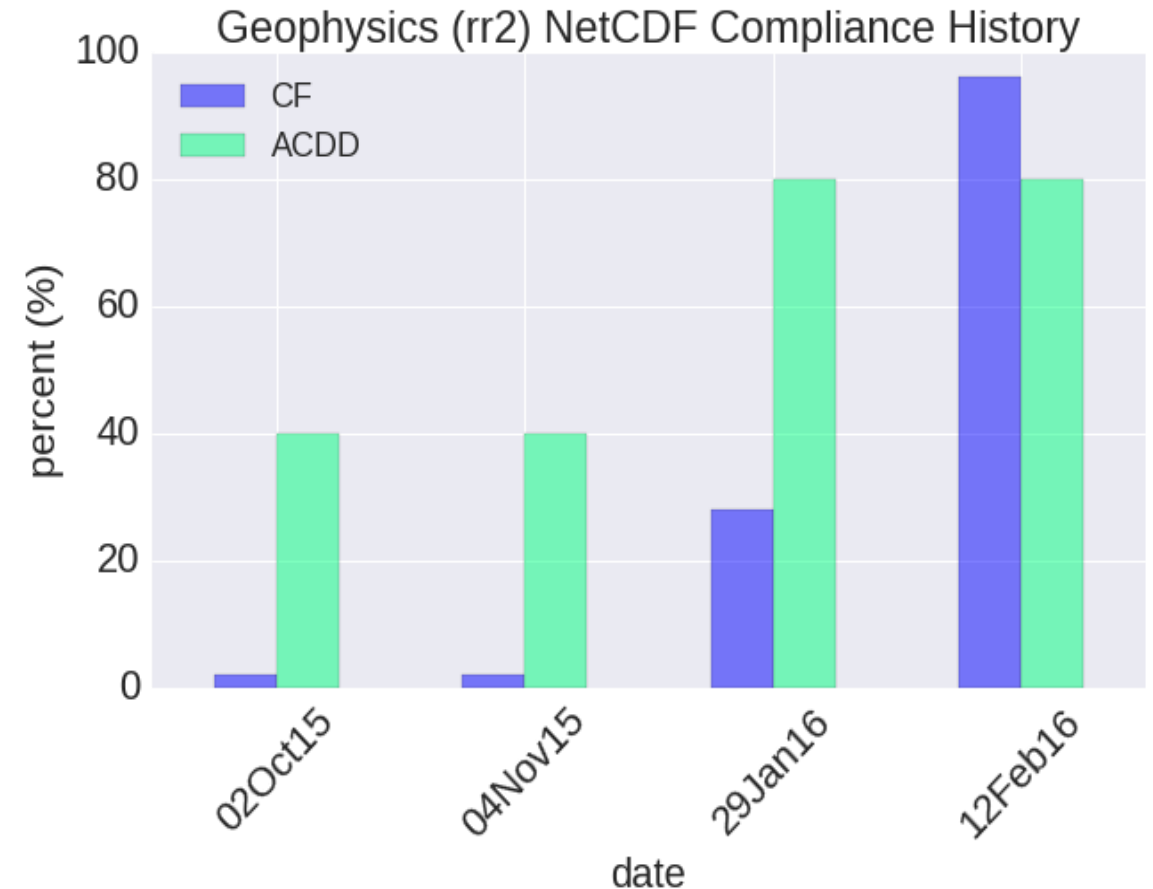


nci.org.au



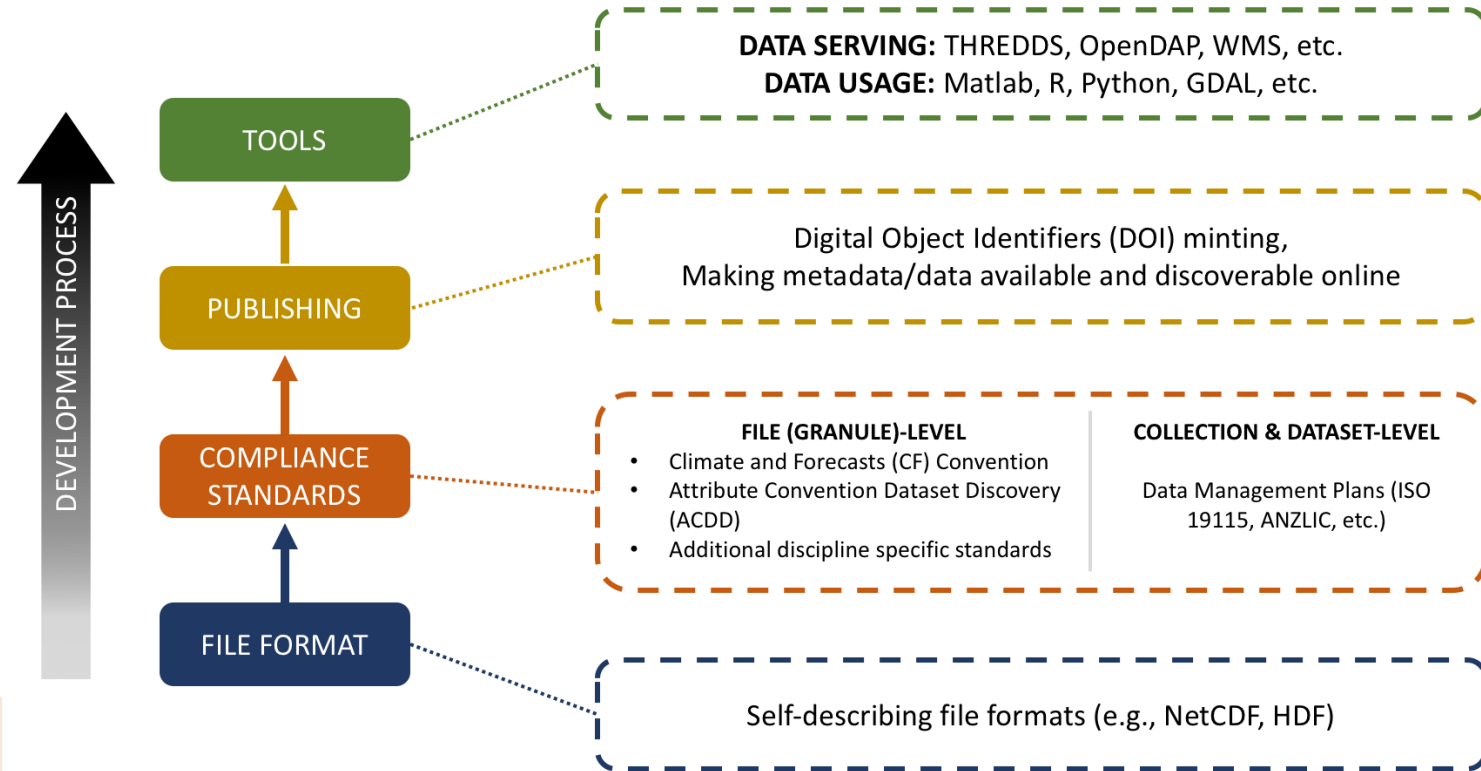
Data Quality Strategy In Action

- Progressive improvement in the quality of the data across the different subject domains
- Improves the ease by which users can access, utilise and combine the datasets from across NCI's holdings



Data Quality Strategy (DQS): What does it involve?

1. Underlying High Performance Data (HPD) file format
2. Close collaboration with data custodians and managers
 - Planning, designing, and assessing the data collections
3. Quality control through compliance with recognised community standards
4. Data assurance through demonstrated functionality across common platforms, tools, and services



- Extend to test “usability” across wide spectrum of scientific tools and data services
 - Commonly used libraries (e.g., netCDF, HDF, GDAL, etc.)
 - Accessibility by data servers (e.g., THREDDS, Hyrax, GeoServer)
 - Validation against scientific analysis and programming platforms (e.g., Python, Matlab, R, QGIS)
 - Visualization tools (e.g., ParaView, IDV, WMS-viewers)



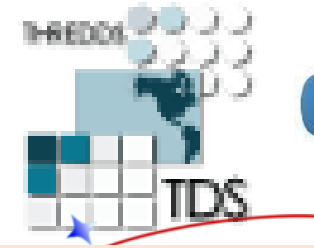
Program/Service	Test	File 1	File 2	File 3	Comments
NetCDF Utilities	ncdump (v4.3.3.1) Read netCDF file contents.	✓	✓	✓	
	NCO (v4.5.3) Read netCDF file contents.	✓	✓	✓	
GDAL Utilities (v1.11.1)	gdalinfo-1 Read netCDF file contents.	✓	✓	✓	
	gdalinfo-2 Read netCDF CRS information.	✓	✓	✓	
Data Viewers	ncview (v2.1.1) Visually inspect netCDF contents.				
	Panoply (v4.5.1) Read and plot netCDF file contents.				
THREDDS Data Server (v4.6)	File download				
	OPeNDAP (access and subsetting) Read/extract netCDF file contents.				
	Netcdf Subset Service (NCSS) Request subset of netCDF contents using spatial/temporal query.				
	Godiva WMS Viewer View netCDF file contents.				
	WMS GetMap (v1.1.1) Request netCDF file using WMS.				
	WCS GetCoverage (v1.0.0) Request netCDF file using WCS.				

Primary motivation:

Positive experience for our users.

Expectation that advertised collections and services are usable.

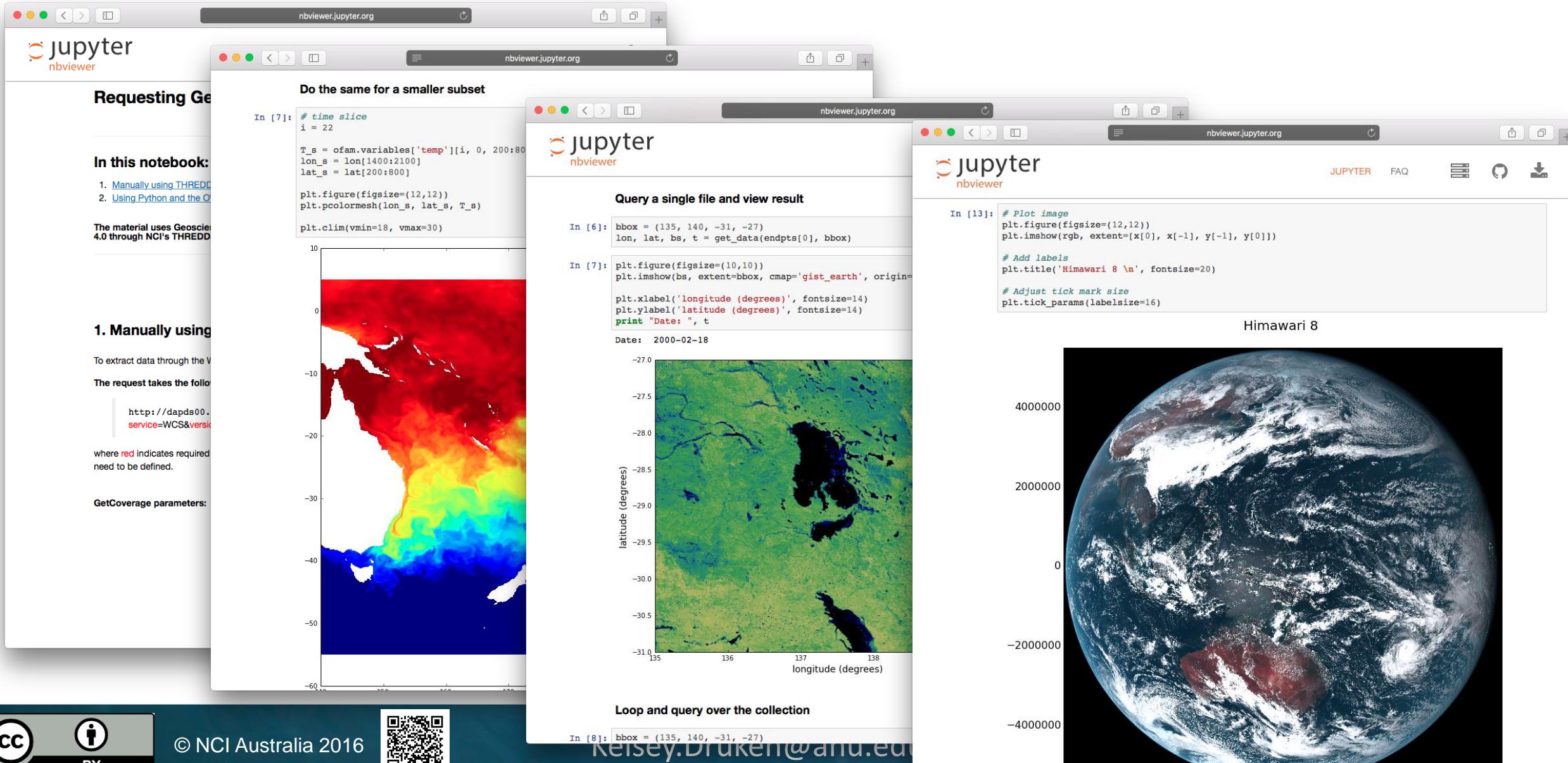
							Read/extract netCDF file contents.				
R (v3.1.0)	ncdf4 (v1.15) Read/extract netCDF file contents.	✓	✓	✓							
QGIS (v2.2.0 Valmiera)	Add data from netCDF as raster layer	N/A	N/A	N/A							
	Add data as WMS layer (served by THREDDS)	N/A	N/A	N/A							
Visualisation Tools	ParaView** (v5.0.1) Read/view netCDF file	N/A	N/A	N/A							



Bonus results:

- Feedback to the local and international communities
→ The more we test and test, the more we learn
- Functionality tests lead to reference and training material for our user community





Requesting Geoscientific Data

In this notebook:

- Manually using THREDDS
- Using Python and the OGC

The material uses Geoscientific Data Discovery (GDD) 4.0 through NCI's THREDDS

1. Manually using THREDDS

To extract data through the VNCI service, the request takes the following form:

```
http://dapds00.nsl.gov.au/geoserver/WCS&version=1.0.0
```

where red indicates required parameters that need to be defined.

GetCoverage parameters:

Do the same for a smaller subset

```
In [7]: # time slice
i = 22

T_s = ofam.variables['temp'][i, 0, 200:800]
lon_s = lon[1400:2100]
lat_s = lat[200:800]

plt.figure(figsize=(12,12))
plt.pcolormesh(lon_s, lat_s, T_s)

plt.clim(vmin=18, vmax=30)
```

Query a single file and view result

```
In [6]: bbox = (135, 140, -31, -27)
lon, lat, bs, t = get_data(endpts[0], bbox)

In [7]: plt.figure(figsize=(10,10))
plt.imshow(bs, extent=bbox, cmap='gist_earth', origin='upper')

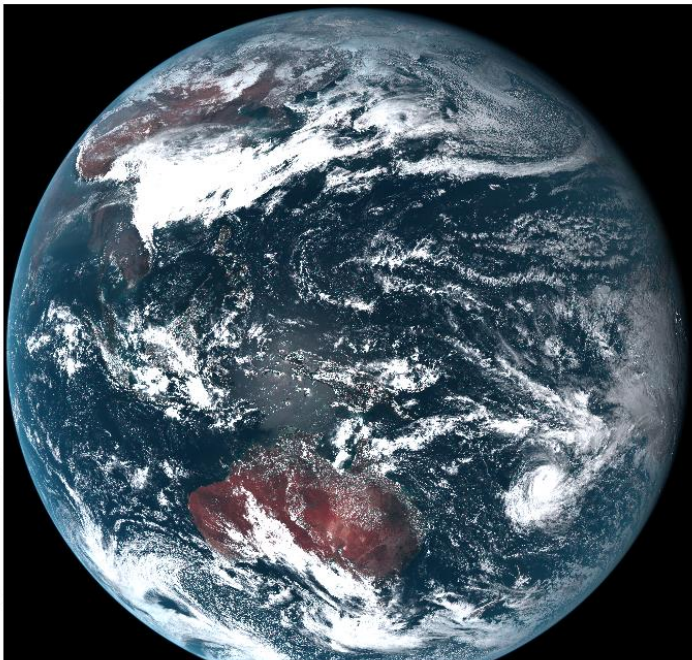
plt.xlabel('longitude (degrees)', fontsize=14)
plt.ylabel('latitude (degrees)', fontsize=14)
print "Date: ", t

Date: 2000-02-18
```

Loop and query over the collection

```
In [8]: bbox = (135, 140, -31, -27)
```

Himawari 8




Bonus results:

- Feedback to the local and international communities
→ The more we test and test, the more we learn
- Functionality tests lead to reference and training material for our user community
- Benefits of standardised and interoperable data formats



What's next?

- Automating and extending these measures and tests across our full collection
- What about the broader file formats?
- Staying connected and working with international communities
 - *E.g., NSF Funded “Advancing netCDF-CF for the Geoscience Community” (EarthCube)*

<https://www.earthcube.org/content/advancing-netcdf-cf-geoscience-community>

