



NCI
AUSTRALIA

A FAIR Data Platform to Support the Next Generation of Transdisciplinary Research at NCI

Lesley Wyborn¹, Ben Evans¹, Clare Richards¹, and
Carina Wyborn²

¹National Computational Infrastructure ANU

²Luc Hoffmann Institute, World Wildlife Fund, University of Montana, Montana, USA

Who the heck is Carina Wyborn?



@rini_rants

- A social scientist with a background in human ecology.
- Her research focuses on: knowledge co-production in climate adaptation and biodiversity conservation and the theory and practice of transdisciplinary, interdisciplinary and integrative research.

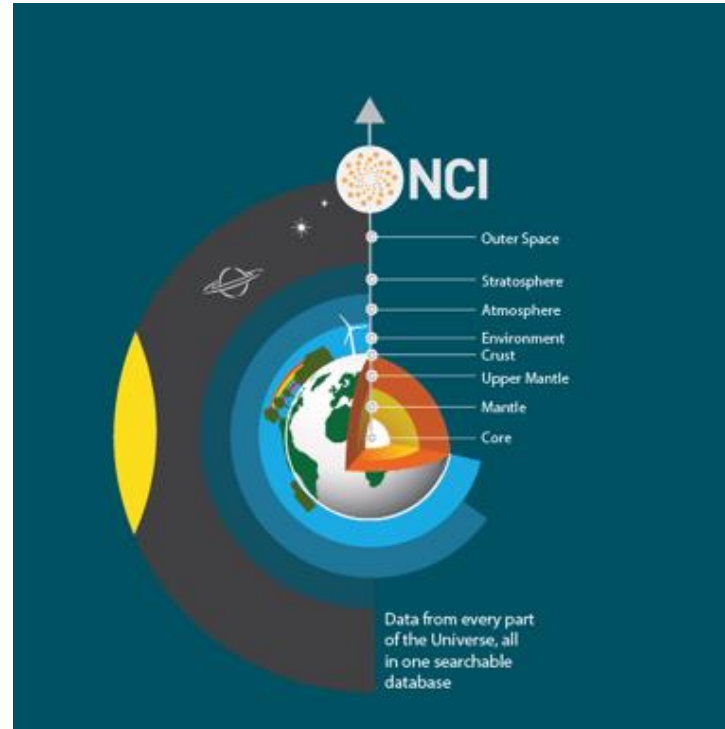
Conclusion: The social scientists are on top of a systematic differentiation between the terms multidisciplinary, interdisciplinary and transdisciplinary

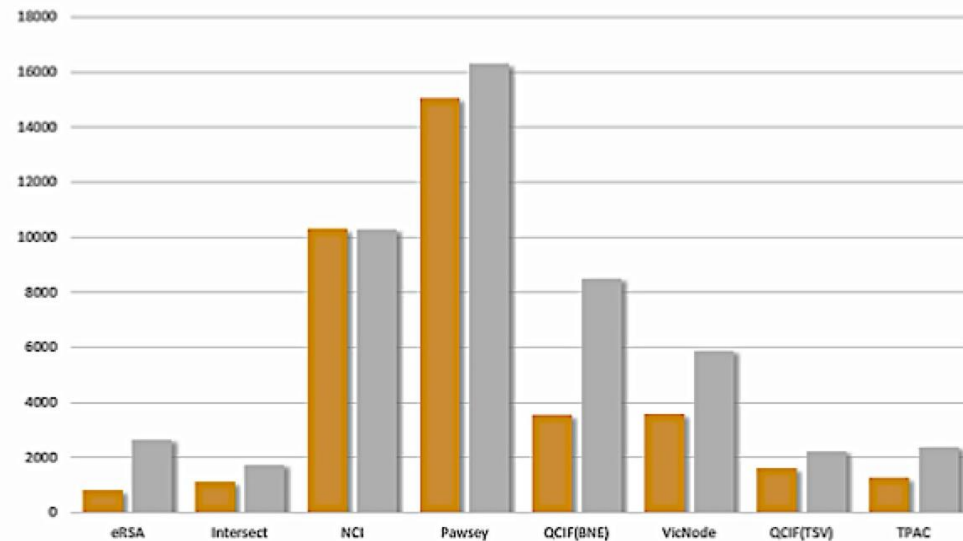
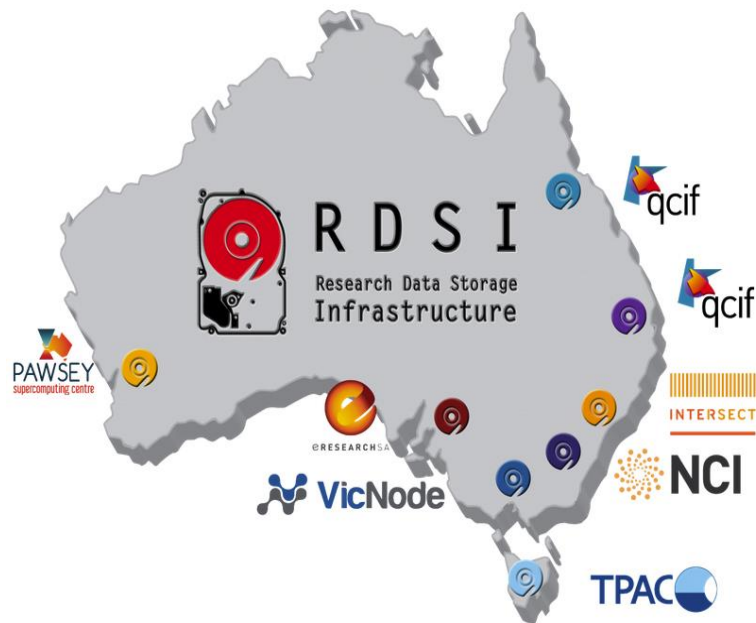
NCI manages 10+PB of reference data collections:

- Climate and Weather, Environmental, Earth Observation, Geophysical, Marine
- Genomics, Optical Astronomy and Social Sciences reference datasets.

Bringing together collections from a range of disciplines – particularly those that naturally interplay across domains.

Locating data within a high performance infrastructure





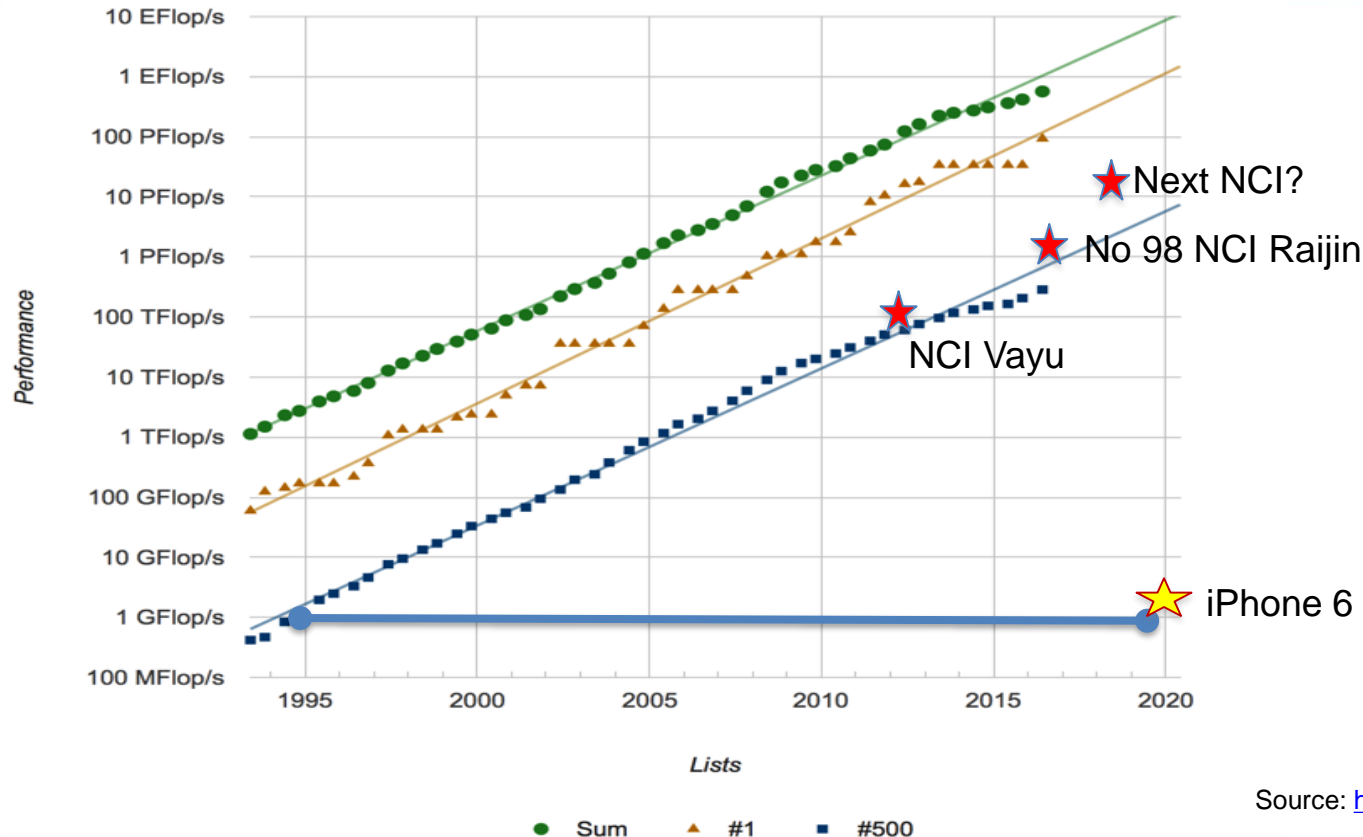
eRSA	Intersect	NCI	Pawsey	QCIF(BNE)	VicNode	QCIF(TSV)	TPAC
821	1132	10296	15036	3531	3575	1596	1267
2679	1765	10296	16322	8520	5884	2270	2405



Total: TB Ready	37254	Total: TB Approved	50141
-----------------	-------	--------------------	-------



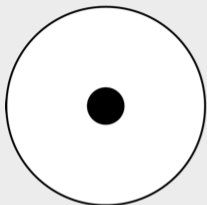
Goal No 2: Our data platform must scale to last the next 10 years



Projected
Performance for
Top 500 HPC

Source: <http://www.top500.org/statistics/perfdevel/>

The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



Intradisciplinary

Working within a single discipline: little attention is paid to cross domain standards



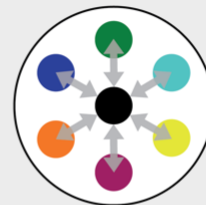
Multidisciplinary

People from different discipline silos working together, but not integrating at the data level



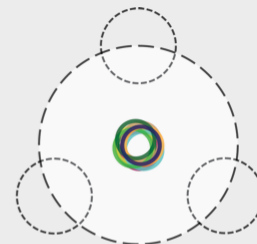
Cross-disciplinary

Data integrated by all disciplines reformatting or interfacing to agreed standards



Interdisciplinary

Data integrated from different disciplines by using brokers that cross walk between the different silos

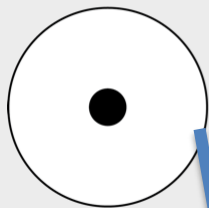


Transdisciplinary

Data is born connected across the discipline boundaries and beyond academia to address societal needs

Trans What?????

The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



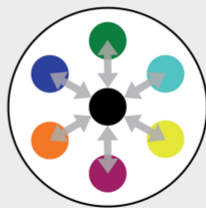
Intradisciplinary



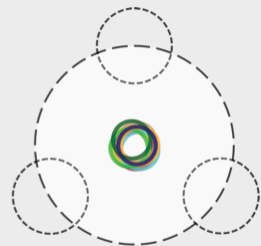
Multidisciplinary



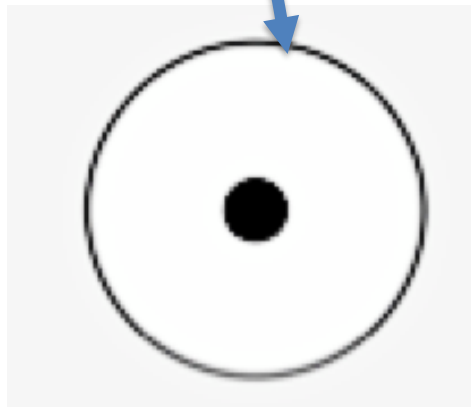
Cross-disciplinary



Interdisciplinary



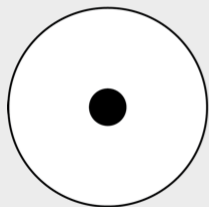
Transdisciplinary



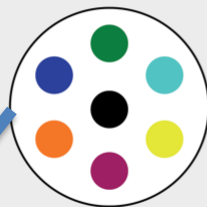
Intradisciplinary

Researchers work within a single discipline or data silo with all participants using the same standard and hence no reformatting or translation of data is required

The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



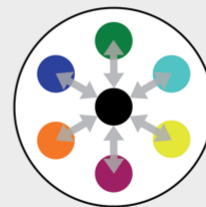
Intradisciplinary



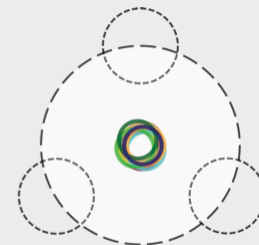
Multidisciplinary



Cross-disciplinary



Interdisciplinary



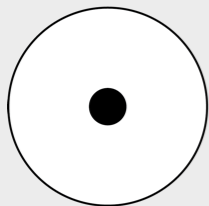
Transdisciplinary



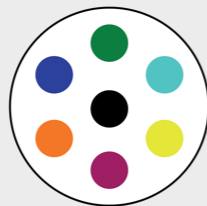
Multidisciplinary

Researchers from different discipline silos work together and share knowledge and results, but are not actually integrating at the data level: outputs are combined at the research paper/report level

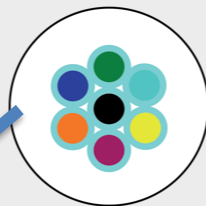
The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



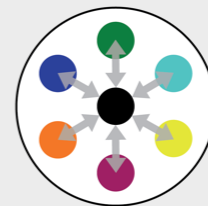
Intradisciplinary



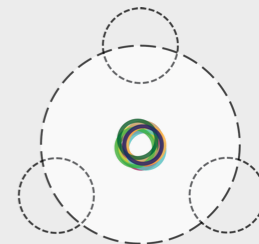
Multidisciplinary



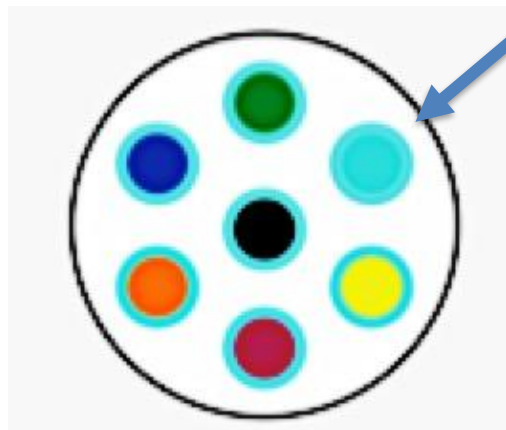
Cross-disciplinary



Interdisciplinary



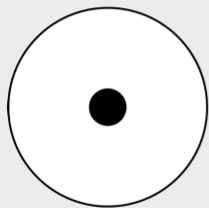
Transdisciplinary



Cross-disciplinary

Researchers participating on a project to integrate data across the groups decide to reformat their datasets to a single agreed suite of specific standards and formats

The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



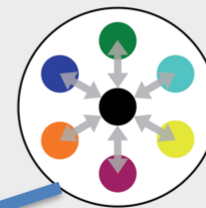
Intradisciplinary



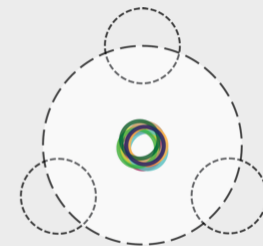
Multidisciplinary



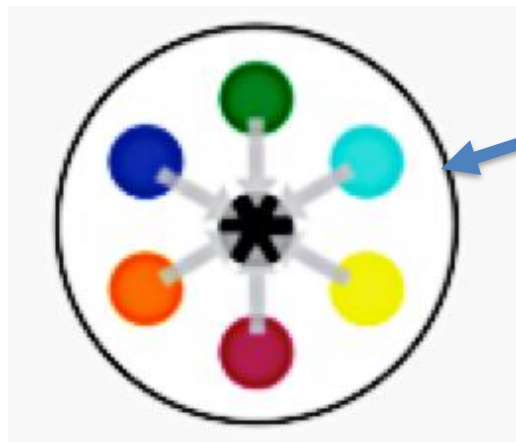
Cross-disciplinary



Interdisciplinary



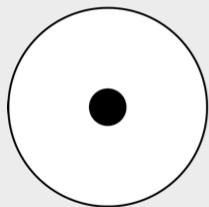
Transdisciplinary



Interdisciplinary

Researchers from each domain integrate their data using customized brokers that cross walk between the different domain silos:
the data of each participant remains unchanged in the back-end

The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



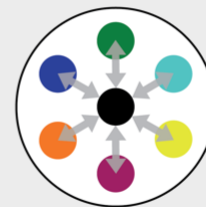
Intradisciplinary



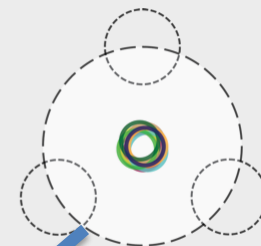
Multidisciplinary



Cross-disciplinary

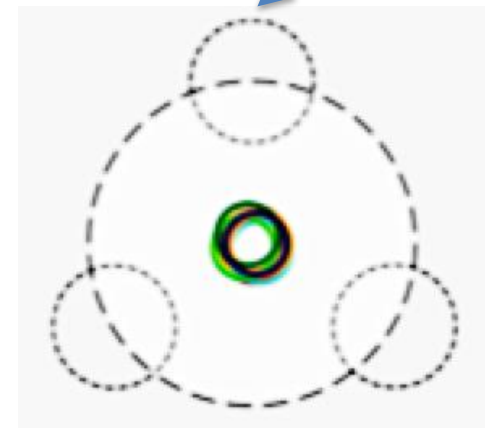


Interdisciplinary



Transdisciplinary

Data is **born connected** to international standards that enable online interaction across the discipline boundaries and beyond academia: researchers participate with stakeholders who can also contribute data



Transdisciplinary

Definitions by the social scientists:

- “A critical and self-reflective research approach that relates societal with scientific problems; it produces new knowledge by integrating different scientific and extra-scientific insights; its aim is to contribute to both societal and scientific progress”

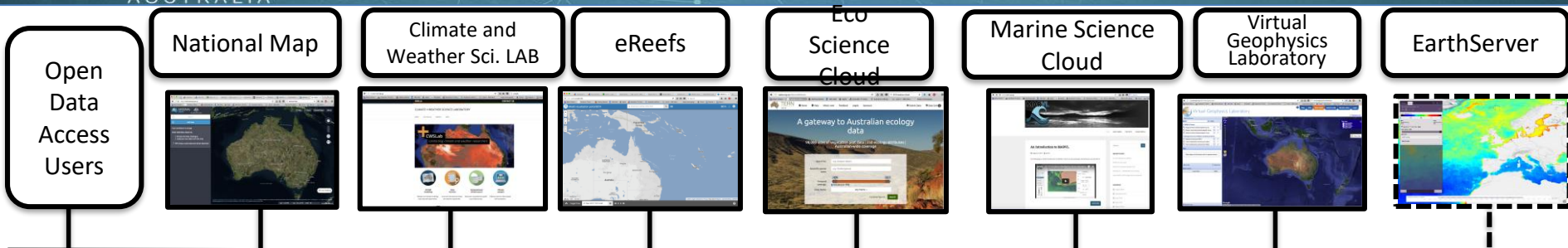
Jahn et al., 2012. Ecological Economics, 79, 1-10.

- “The rationale for transdisciplinarity is global challenges, which are complex”

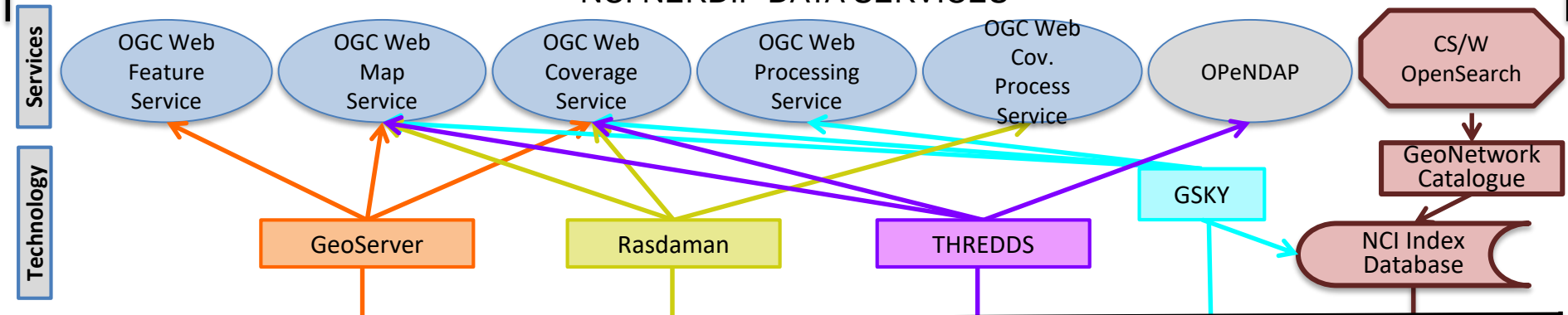
Vanasupa et al., 2014. Sustainability, 6, 2893-2928

Definition by the informaticians

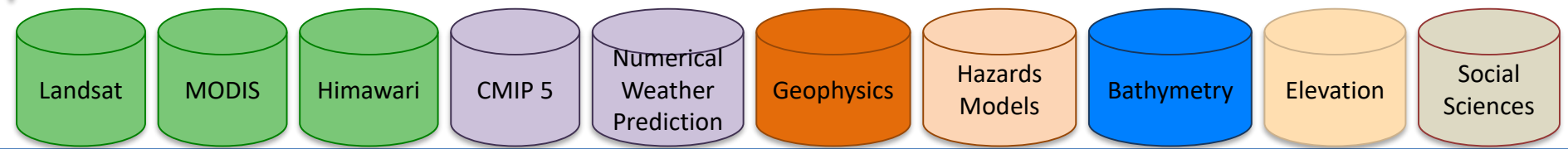
- Researchers across the science disciplines, the humanities, the social sciences and those beyond academia need to work together to create integrated data platforms that interoperate horizontally across discipline boundaries, and enable access to data by a diversity of users from high end researchers, to undergraduates and to the general public.



NCI NERDIP DATA SERVICES

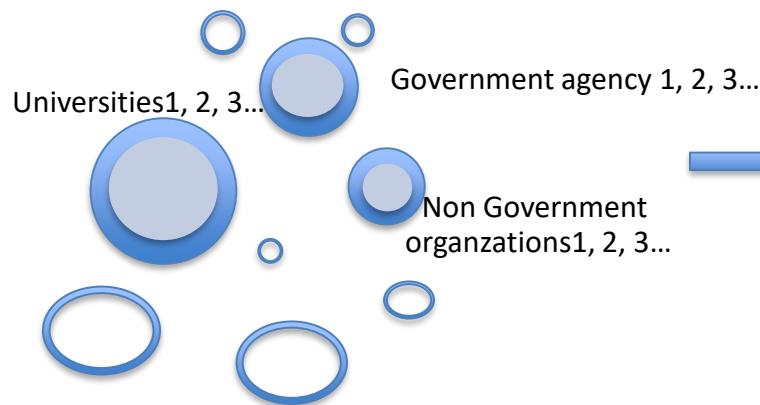


10 PB NCI NERDIP EARTH SYSTEMS, ENVIROMENTAL AND SOLID EARTH DATA COLLECTIONS

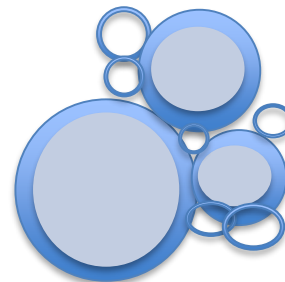




Disparate data collections

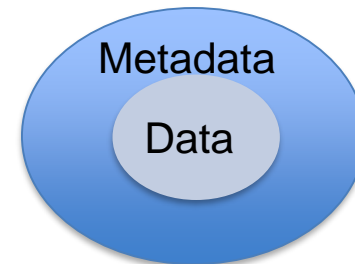


Curated data collections



step 1


Ready for transdisciplinary data access, including services



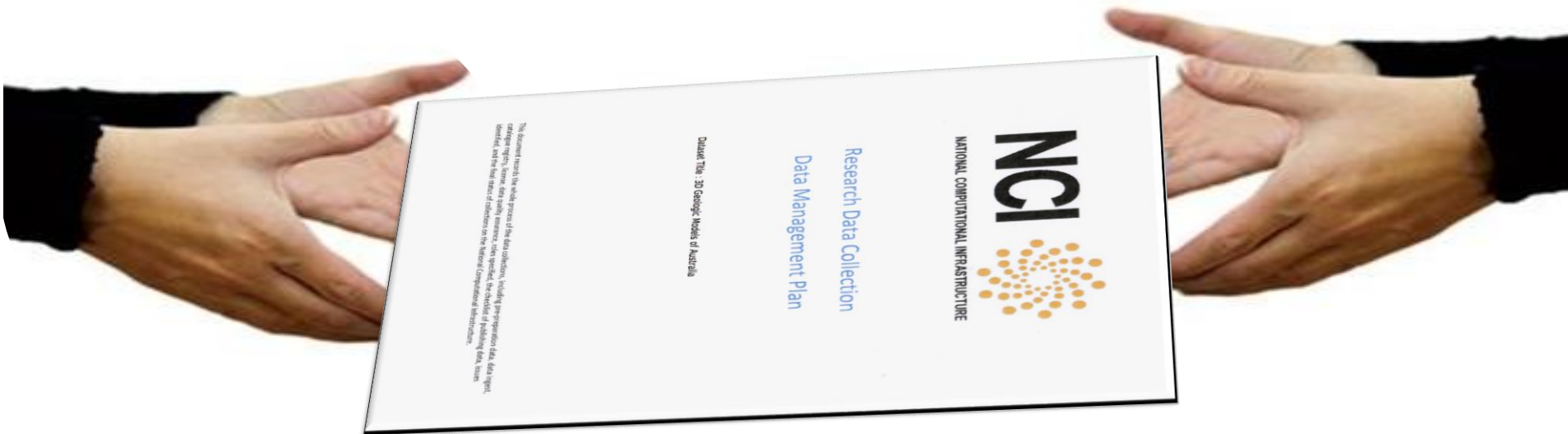
step 2



Mutually agreed plan on how the collection will be managed and published



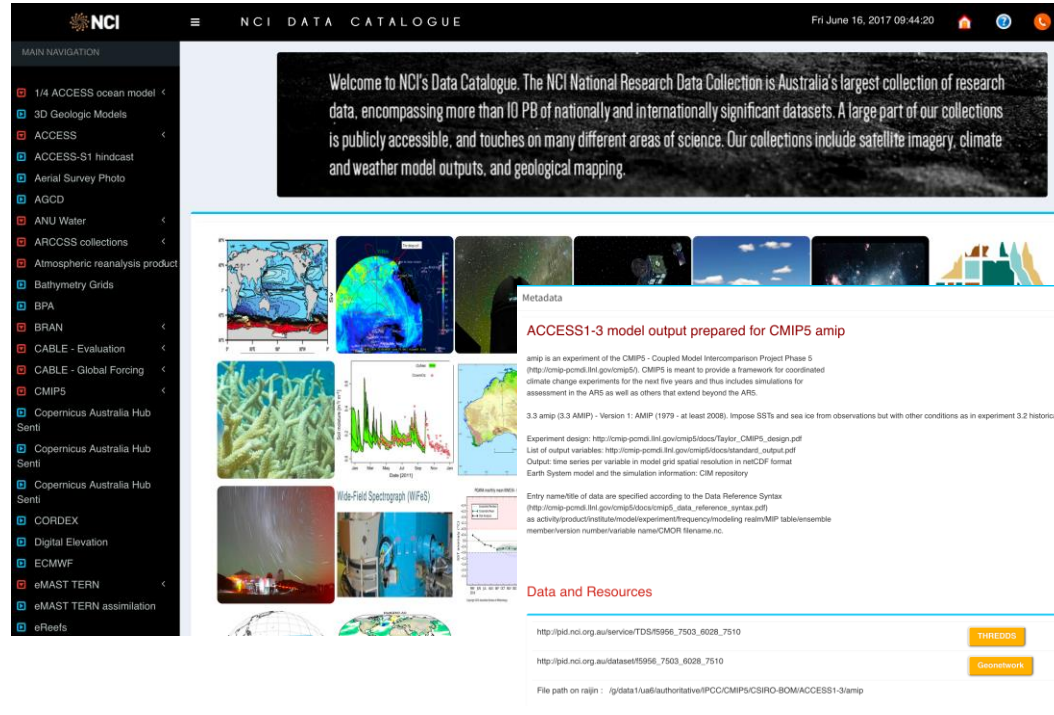
NCI Data Manager



Source: http://www.moneymarketing.co.uk/pictures/620xAny/9/1/3/2080913_Business-Handover-Finance-General-700.jpg

Findable:

- Datasets and catalogue entries are both human readable and machine harvestable
- Can cross-walk metadata with ISO 19115, RIF-CS and DCAT
- Findable Research Data Australia, NCI and custodians catalogues
- International discoverability – hosting and federating with international collections (compliance with standards)
- Open access to discover the data: a small number of cases may require permission to access it



The screenshot shows the NCI Data Catalogue interface. The top navigation bar includes the NCI logo, a menu icon, the text "NCI DATA CATALOGUE", and the date "Fri June 16, 2017 09:44:20". A large banner at the top reads: "Welcome to NCI's Data Catalogue. The NCI National Research Data Collection is Australia's largest collection of research data, encompassing more than 10 PB of nationally and internationally significant datasets. A large part of our collections is publicly accessible, and touches on many different areas of science. Our collections include satellite imagery, climate and weather model outputs, and geological mapping."

On the left is a "MAIN NAVIGATION" sidebar with a list of categories and their counts: 1/4 ACCESS ocean model, 3D Geologic Models, ACCESS, ACCESS-S1 hindcast, Aerial Survey Photo, AGCD, ANU Water, ARCCSS collections, Atmospheric reanalysis product, Bathymetry Grids, BPA, BRAN, CABLE - Evaluation, CABLE - Global Forcing, CMIP5, Copernicus Australia Hub Senti, Copernicus Australia Hub Senti, Copernicus Australia Hub Senti, CORDEX, Digital Elevation, ECMWF, eMAST TERN, eMAST TERN assimilation, and eReefs.

The main content area displays a grid of data visualizations, including maps, charts, and satellite imagery. A "Metadata" pop-up window is visible, titled "ACCESS1-3 model output prepared for CMIP5 amip". It contains the following text:

ACCESS1-3 model output prepared for CMIP5 amip

amip is an experiment of the CMIP5 - Coupled Model Intercomparison Project Phase 5 (<http://comp-pcm5.lrl.gov/cmip5/>). CMIP5 is meant to provide a framework for coordinated climate change experiments for the next five years and thus includes simulations for assessment in the AR5 as well as others that extend beyond the AR5.

3.3 amip (3.3 AMIP) - Version 1: AMIP (1979 - at least 2008). Impose SSTs and sea ice from observations but with other conditions as in experiment 3.2 historical.

Experiment design: http://comp-pcm5.lrl.gov/cmip5/docs/Taylor_CMIP5_design.pdf
 List of output variables: http://comp-pcm5.lrl.gov/cmip5/docs/standard_output.pdf
 Output: time series per variable in model grid spatial resolution in netCDF format
 Earth System model and the simulation information: CIM repository

Entry name/tile of data are specified according to the Data Reference Syntax (http://comp-pcm5.lrl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf)
 as activity product/institute/model/experiment/frequency/modeling realm/MIP table/ensemble member/version number/variable name/CMOR term name.nc

Below the metadata window, there is a "Data and Resources" section with the following links:

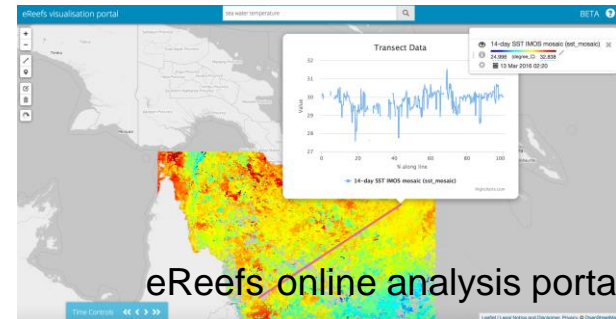
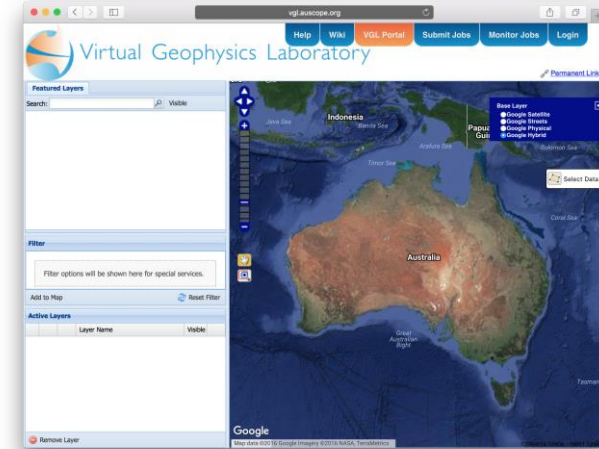
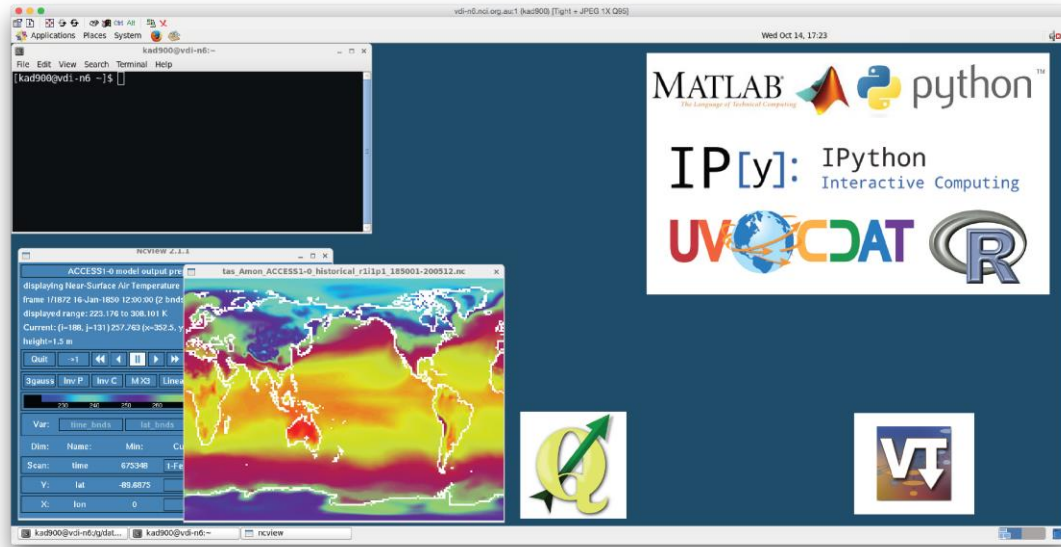
http://pdx.nci.org.au/service/TDS/IS956_7503_6028_7510 [THREDDS](#)

http://pdx.nci.org.au/dataset/IS956_7503_6028_7510 [Download](#)

File path on rajn : /gdata1/au/authoritative/PC/CMIP5/CSIRO-BOM/ACCESS1-3/amip

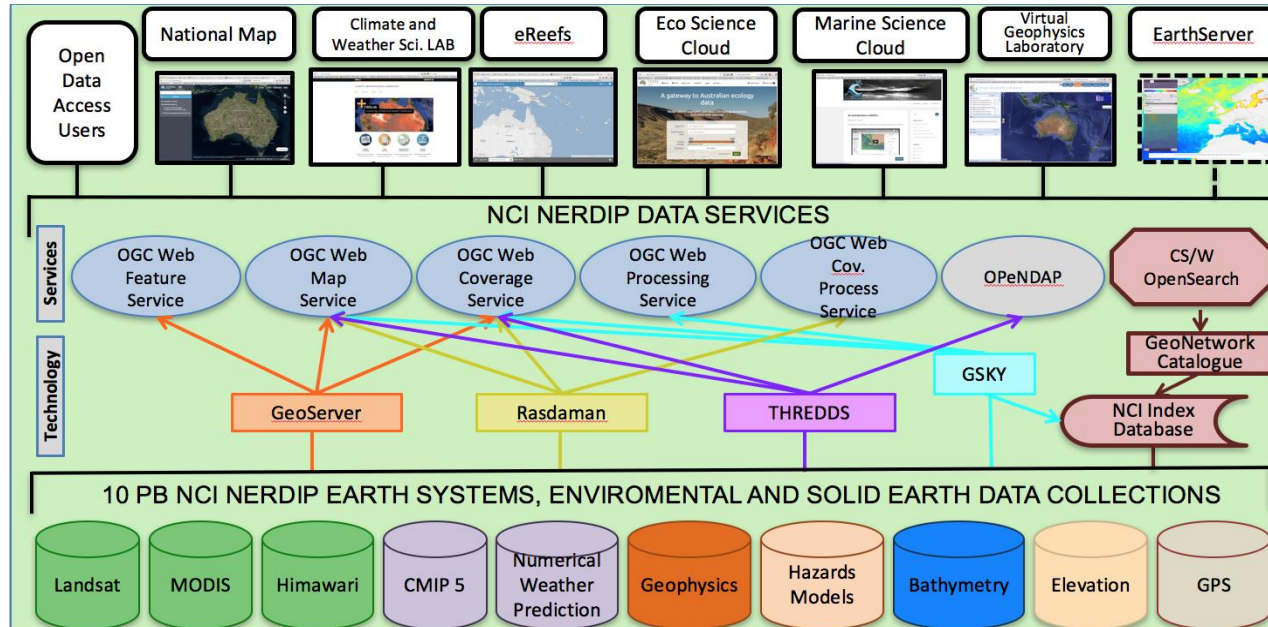
Collections can be accessed from a broad range of options

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



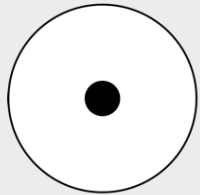
eReefs online analysis portal

- 1) general research data access (data download for small file sizes),
- 2) advanced techniques and portals for multiple applications
 - e.g. Virtual labs, Portals, well-known desktop tools, programmatic access via network or in-situ



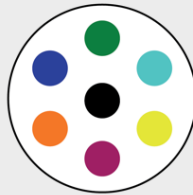
Aim for a transdisciplinary approach for a diverse range of users: leverage common approaches wherever possible

The 'Disciplinary' Data Integration Spectrum: Where do You Sit?



Intradisciplinary

Working within a single discipline: little attention is paid to cross domain standards



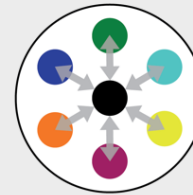
Multidisciplinary

People from different discipline silos working together, but not integrating at the data level



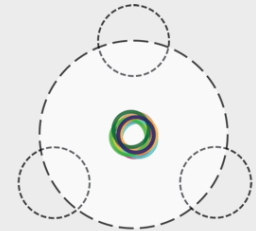
Cross-disciplinary

Data integrated by all disciplines reformatting or interfacing to agreed standards



Interdisciplinary

Data integrated from different disciplines by using brokers that cross walk between the different silos

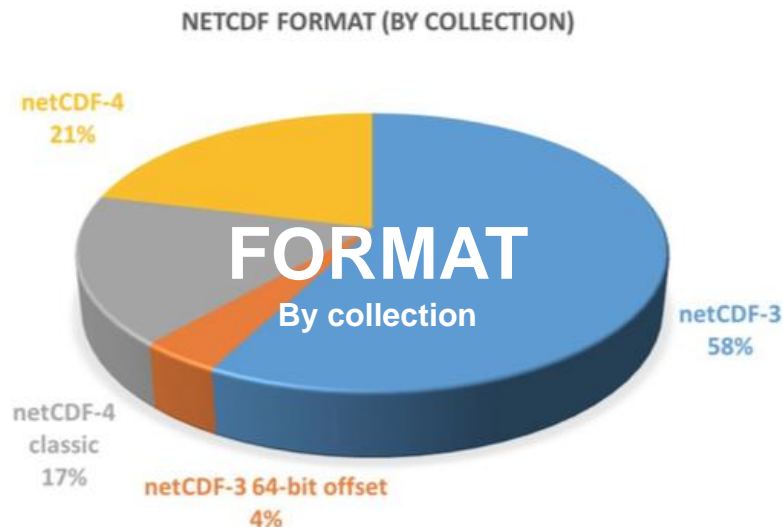
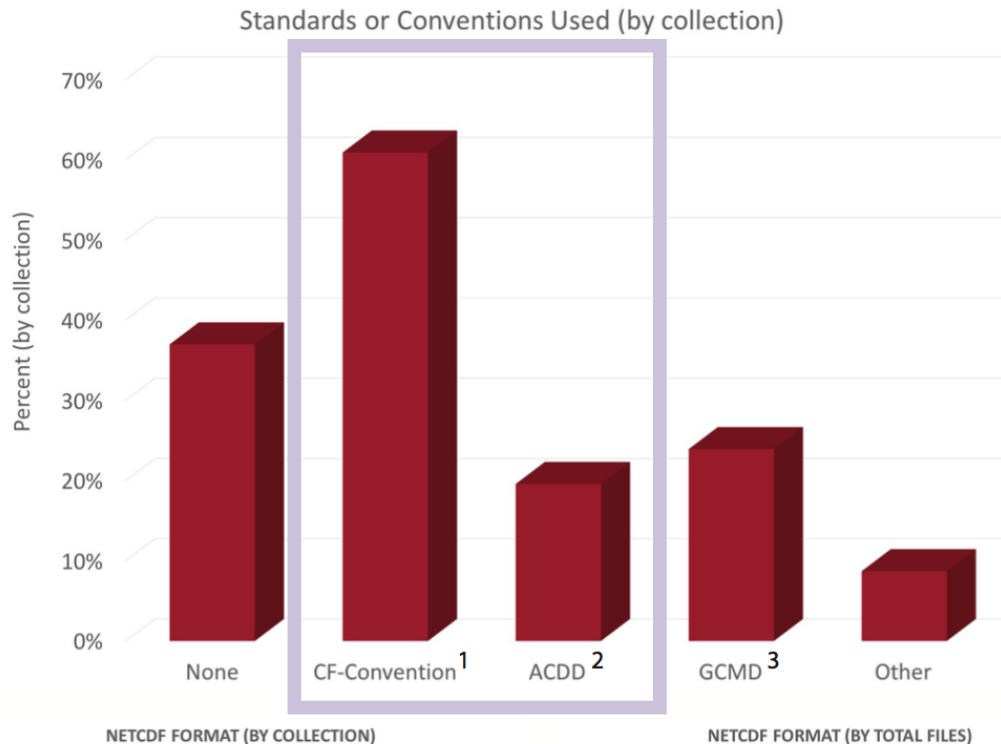


Transdisciplinary

Data is born connected across the discipline boundaries and beyond academia to address societal needs

Applying international data standards for interoperability

- Metadata standards at both data services (e.g. catalogues) and at the data level
- Controlled vocabularies for data
- Interchangeable self-describing formats for data (e.g., netCDF4/HDF5)



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]

Ensuring datasets meet community standards

Summarised version on the compliance status.

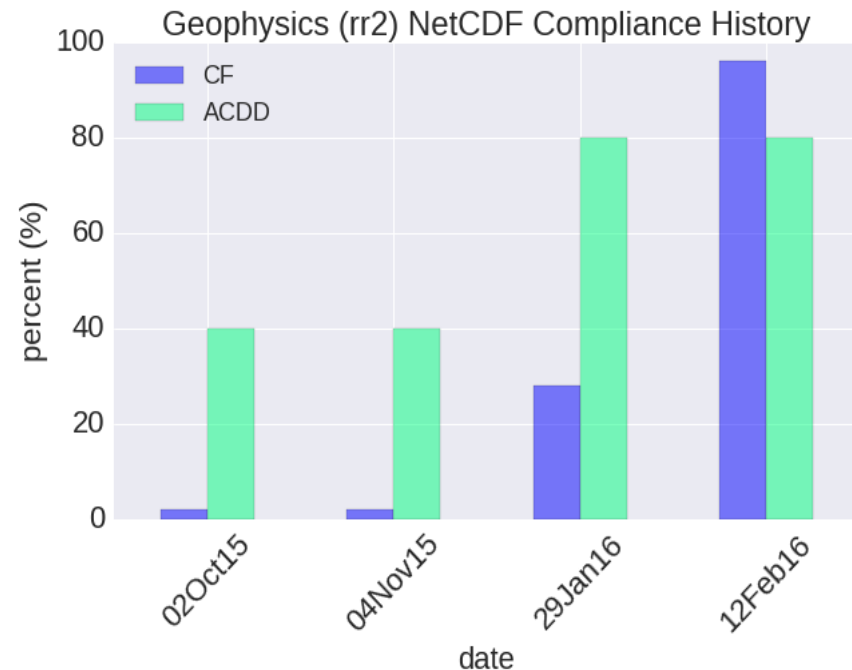
The break down... compliance scores and also measure of consistency across the collection

Providing attack plan for improvements:

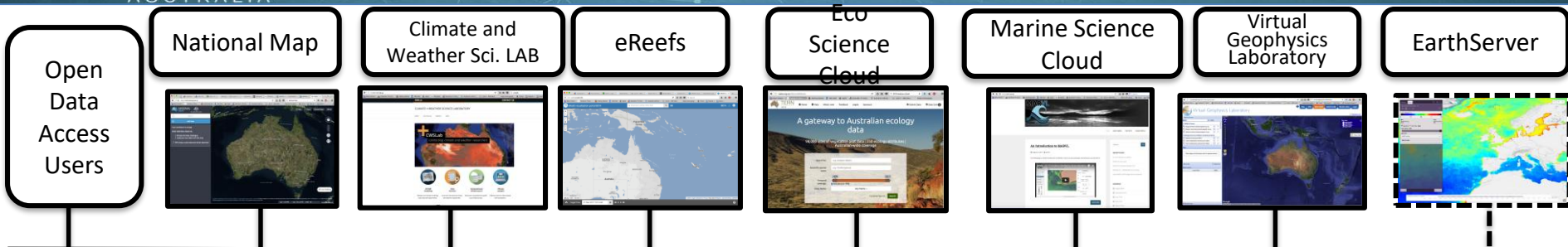
Make it easy for data managers to efficiently address and meet baseline compliance

Data Quality Strategy In Action

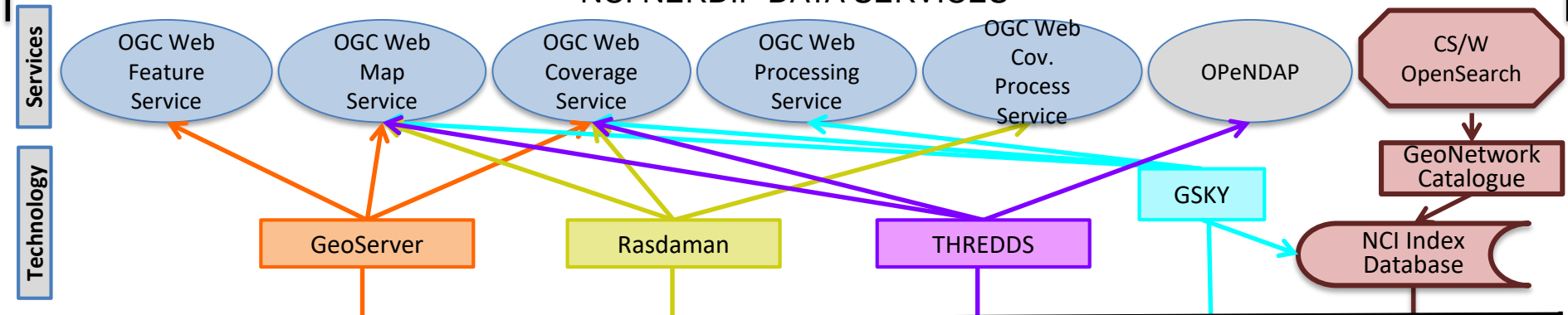
- Progressive improvement in the quality of the data across the different subject domains
- Improves the ease by which users can access, utilise and combine the datasets from across NCI's holdings



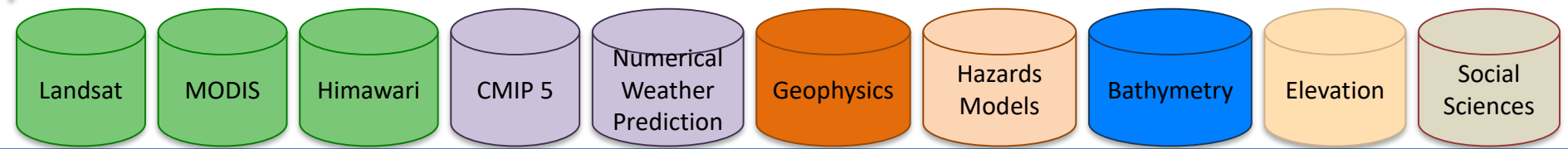
[illegible]

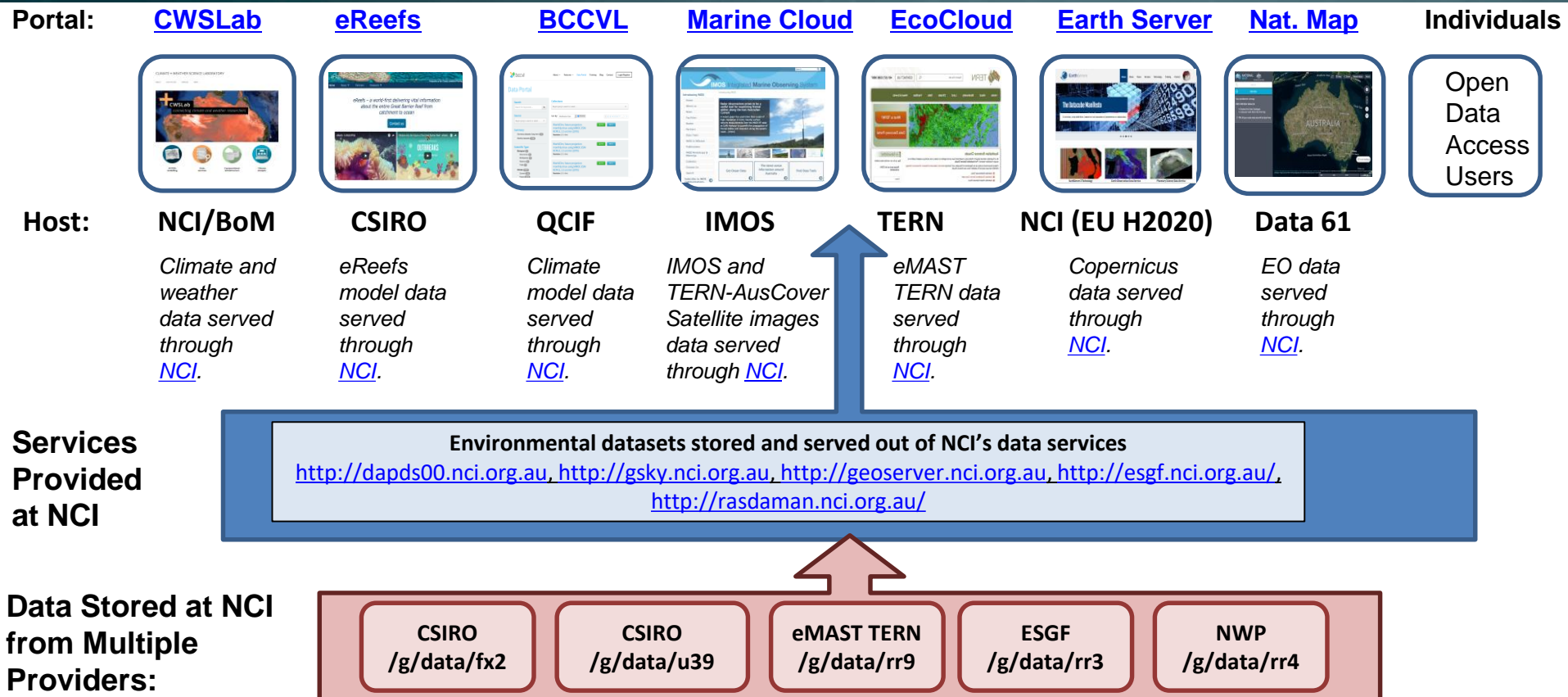


NCI NERDIP DATA SERVICES



10 PB NCI NERDIP EARTH SYSTEMS, ENVIROMENTAL AND SOLID EARTH DATA COLLECTIONS





Portal:

[VGL](#)

[AuScope](#)

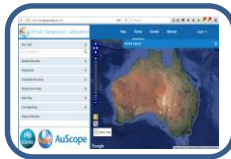
[GA Catalogue](#)

[AusGIN](#)

[ANVGL](#)

[Nat Map](#)

Open
Data
Access
Users



Host Agency:

CSIRO

CSIRO

GA

GA

GA

Data 61

National Coverages and Airborne Geophysics TDS services from [NCI](#); WMS from [GA](#). For some files, WCS and WMS from [NCI dap-wms server](#).

ASTER files from [NCI dap-wms server](#). GSWA Gravity, Mag & Radiometric Grid from [NCI](#)

National Coverages and Airborne Geophysics TDS services from [NCI](#); WMS from [GA](#). For some files, WCS and WMS from [NCI dap-wms server](#).

ASTER files from [NCI dap-wms server](#). National Coverages from [GA](#).

National Coverages from [NCI](#).

Geophysics data served through [NCI](#).

Services
Provided by NCI

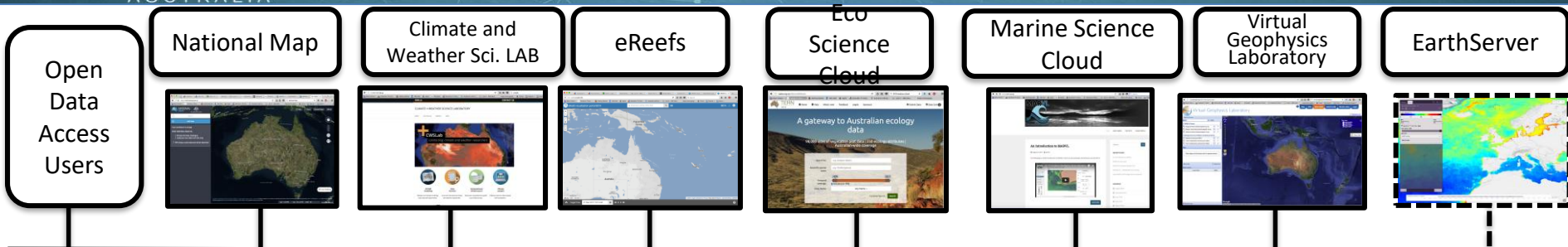
Geophysics datasets stored and served out of NCI's data services
<http://dapds00.nci.org.au> or <http://dap-wms.nci.org.au>

Data Provider:

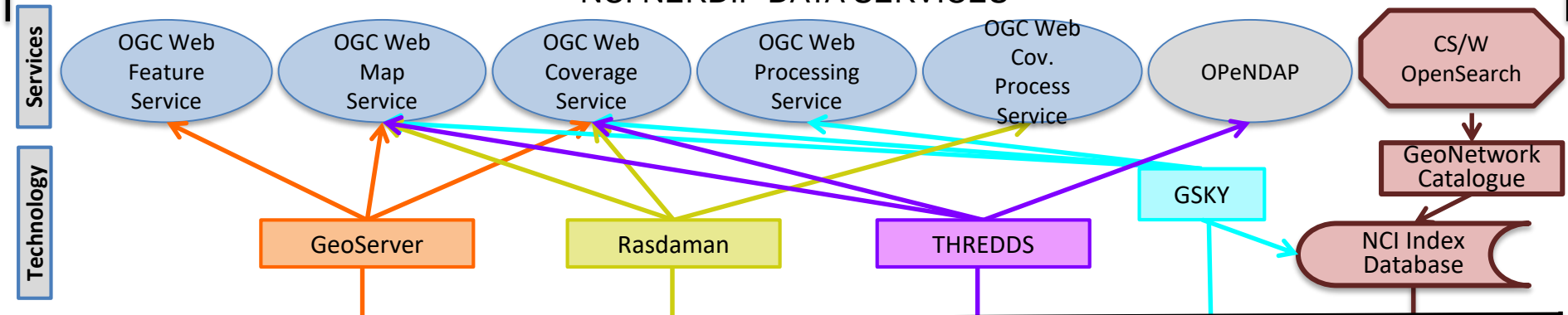
GA /g/data/rr2

GSWA /g/data/rl1

GA /g/data/wx7



NCI NERDIP DATA SERVICES



10 PB NCI NERDIP EARTH SYSTEMS, ENVIROMENTAL AND SOLID EARTH DATA COLLECTIONS

