2019 AMOS Darwin

NCI AUSTRALIA

Introduction to NCI's Data Collections

Jingbo Wang & Kate Snow

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

Australian Government Bureau of Meteorology

Australian Government Geoscience Australia

Australian Government Australian Research Council

CSIRO

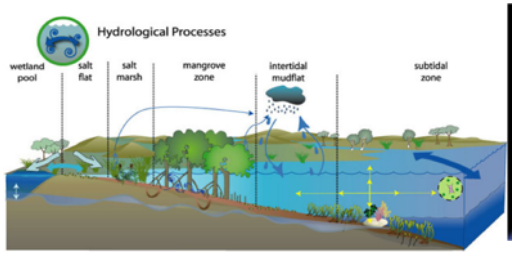
Australian National University

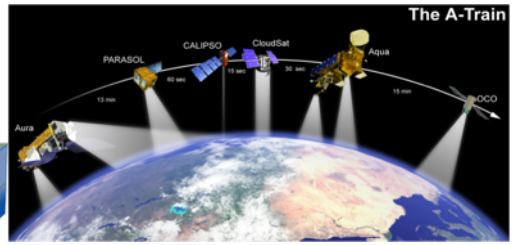
nci.org.au

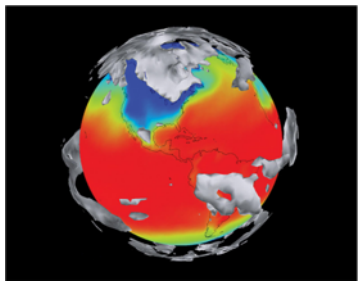
@NCInews

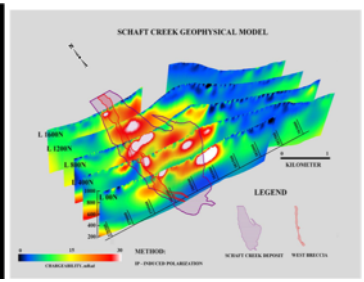
Overview

NCI makes available national reference datasets – especially those produced by the government agencies. It is brought together at NCI and organised for both high performance computation & high performance data analysis, as well as making available more broadly to the research community.

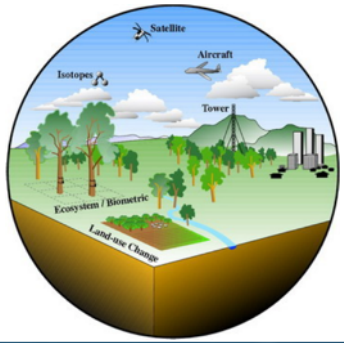








- ***climate and weather models***
- ***satellite images***
- ***bathymetry and elevation***
- ***hydrology***
- ***geophysics***
- ***Also: optical astro, genomic and social sciences***



Overview

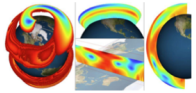
User generate/transfer data

Data Manager fill DMP and create catalogue

Super computer users

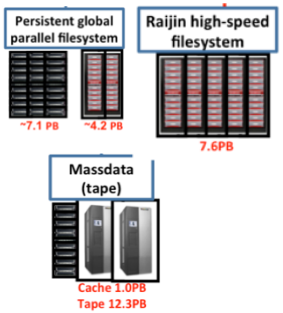
Paper and Data are published

Data visualization



Data share and re-use

NCI provides user with Data-as-a-Service

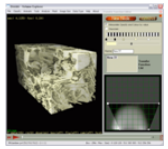


Fast data storage

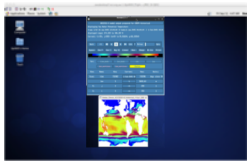
Data Management Portal

HPC

Data Curation, Publish, Citation



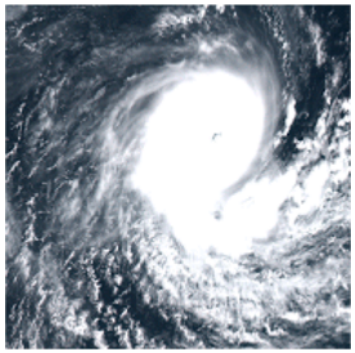
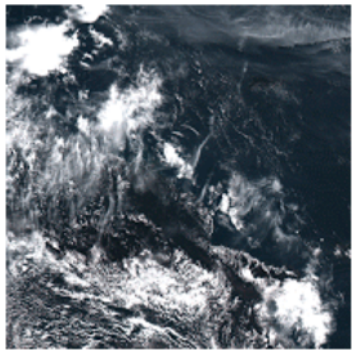
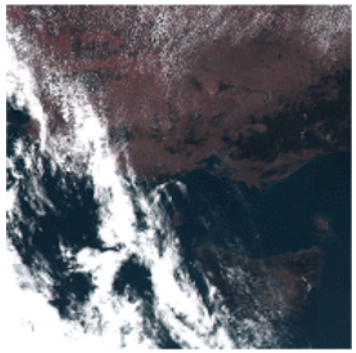
Visualisation tools



Web-time analytics software

Range of data collections

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

Overview

Data Collections	NCI project codes
CMIP6 replicated	oi10
CMIP5 Australian published / replicated	rr3 / al33
CMIP3 entire data collection	cb20
CORDEX	rr3, al33
Input4MIPs include JRA55-do (forcing for ocean/sea-ice models)	qv56
ERA-Interim (6hrly data)	ub4
ACCESS NWP models (BoM)	ja4, lb4, na3
LANDSAT, MODIS, VIIRS, AVHRR, INSAR, MERIS	rs0, fk4, u39
BoM Seasonal Climate	rr8, ub7
BoM Observations - Himawari Satellite	rr5
Ocean-Marine	rr6, gb6
ARCCSS/CLEX collections, including ERA-Interim (6hrly)	ua8, ub4
YOTC, ACCESS-CM, CABLE	rq7, gh5, wd9

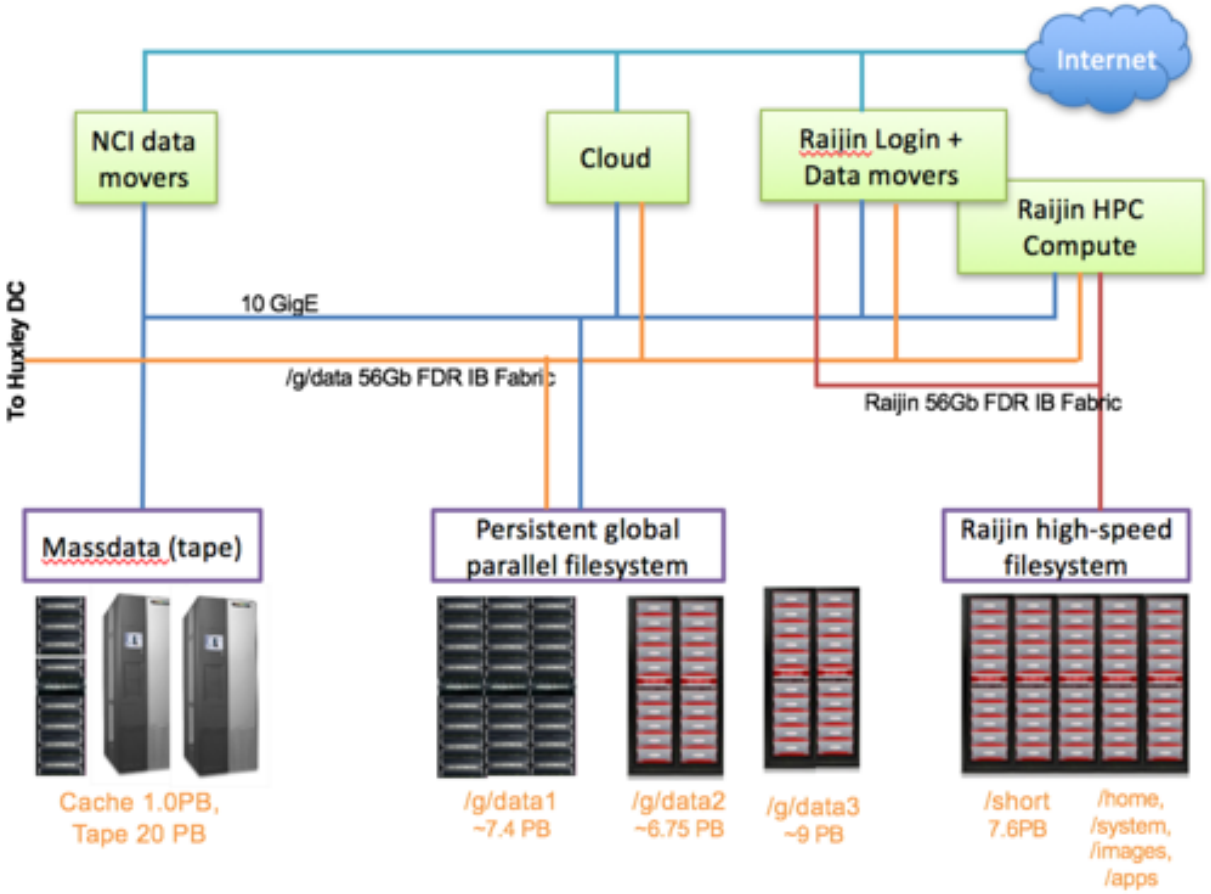
Overview

Datasets are stored on the NCI global filesystems. These are called:

- /g/data1a
- /g/data1b
- /g/data3
- /g/data4

These are all symlinked from **/g/data** (you don't have to remember the filesystem number).

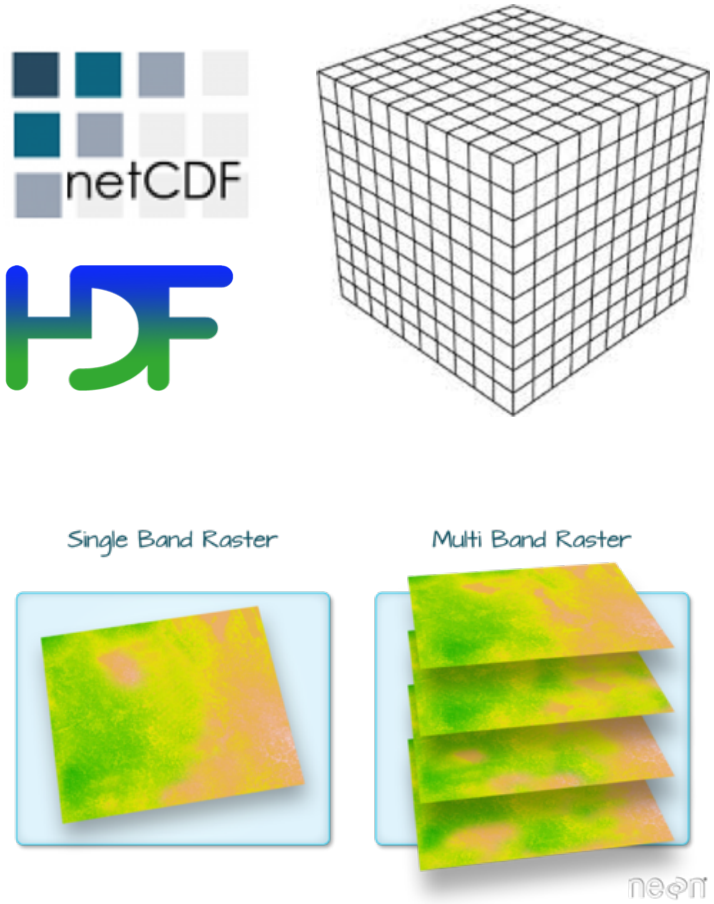
The data stored on these filesystems are available to NCI's HPC, the VDI, and some data services.



Types of data formats

Range of formats:

- NetCDF/HDF5
- GeoTIFF
- GRIB
- ...
- (Proprietary formats)



Network Common Data Form (NetCDF) & Hierarchical Data Format (HDF5)

- Support a wide variety of data types as well as data structures:
  - Scientific data arrays
  - Tables
  - Raster/image data
  - String data
  - Etc...
- Used across large (and growing) spectrum of subject disciplines

  Big pro: Common formats enable transdisciplinary interoperability.

Network Common Data Form (NetCDF) & Hierarchical Data Format (HDF)

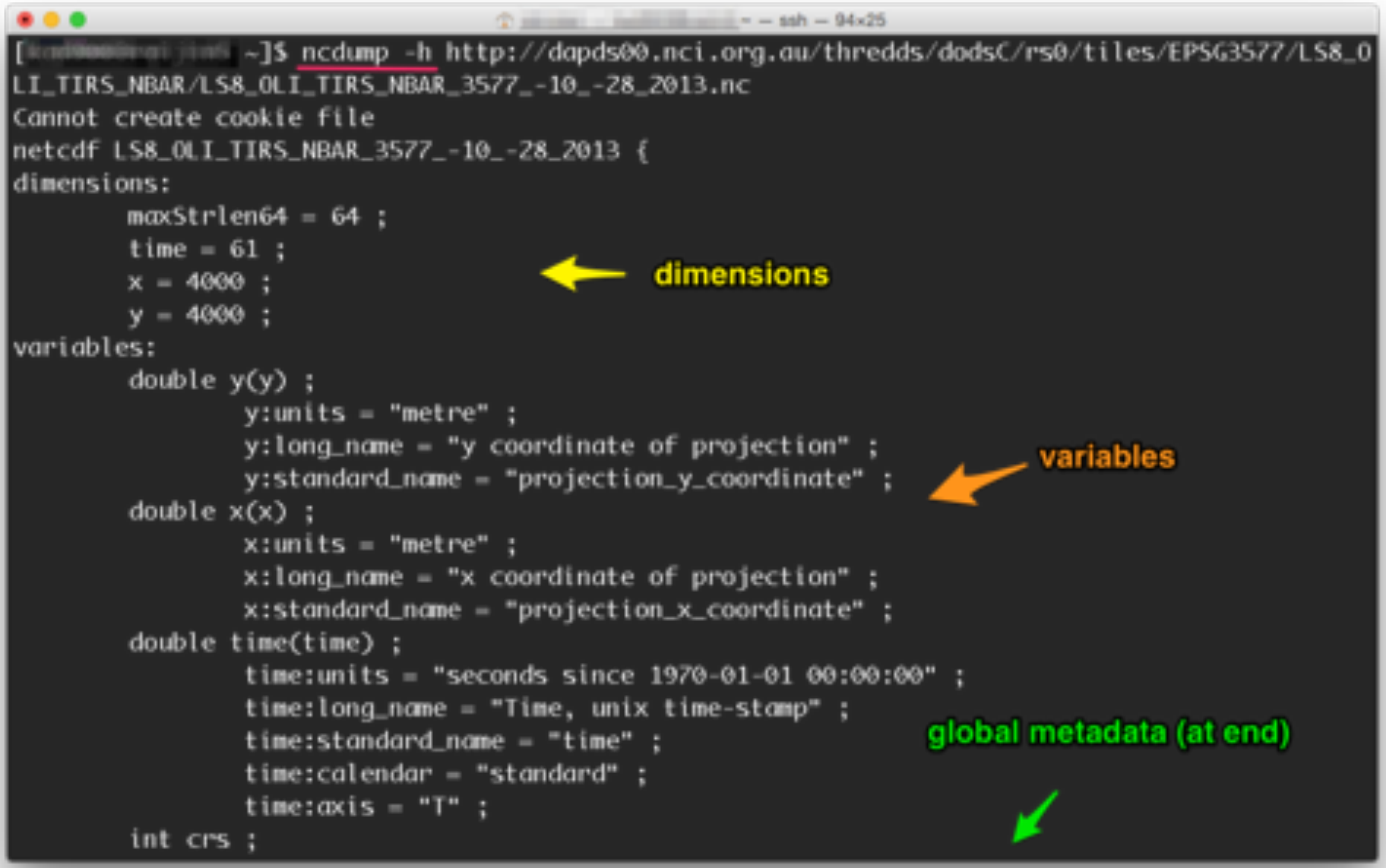
(from Unidata: http://www.unidata.ucar.edu/software/netcdf/docs/faq.html#whatisit)

- **Self-Describing.** A netCDF file includes information about the data it contains.
- **Portable.** A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- **Scalable.** A small subset of a large dataset may be accessed efficiently.
- **Appendable.** Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
- **Sharable.** One writer and multiple readers may simultaneously access the same netCDF file.
- **Archivable.** Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

A quick look inside...

In particular, what does “self-describing” look like?

- File metadata information
- Dimensions
- Variables
- Variable-level metadata
- Special attributes
  - Compression
  - Chunking
  - Endianness



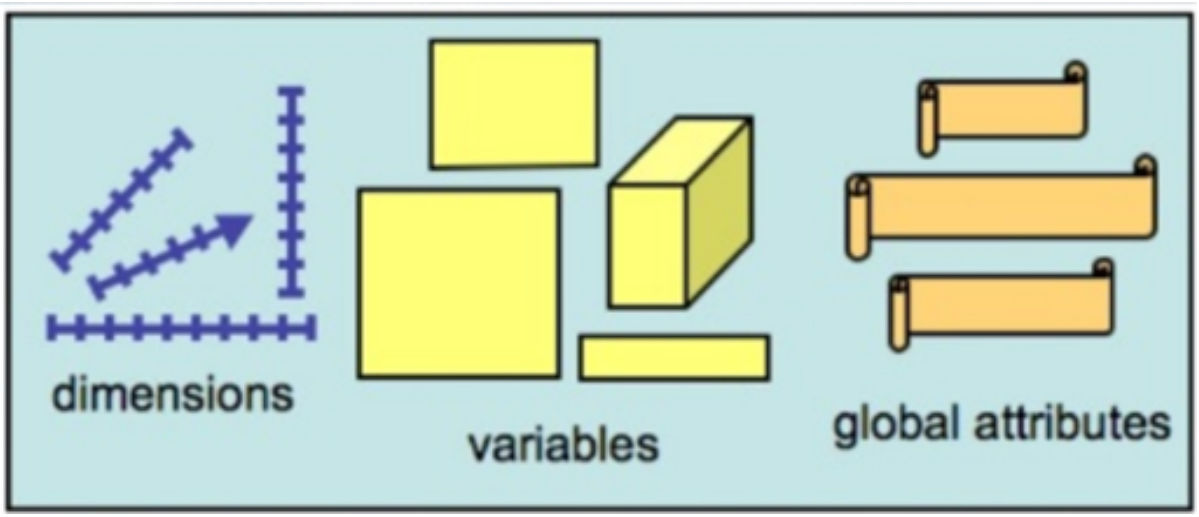
Types of NetCDF files

Classic Data Model:

- ➔ NetCDF3 (or just “classic”)
- ➔ 64-bit offset format

Enhanced Data Model:

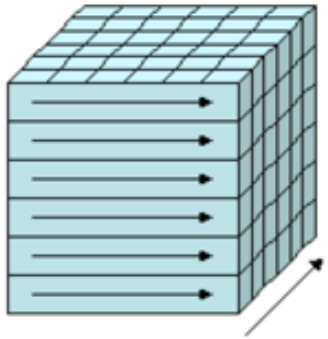
- ➔ NetCDF4
- ➔ NetCDF4-classic



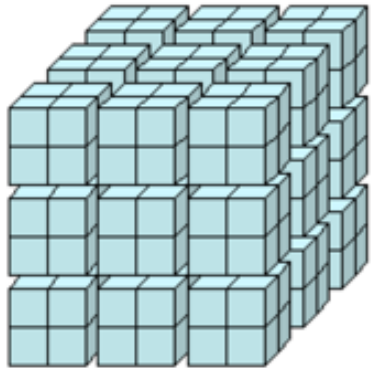
http://www.slideshare.net/HDFEOS/netcdf4-tutorial-ws14

NetCDF/HDF5 Structure Concepts

NetCDF and HDF5 are multidimensional array data containers:



index order



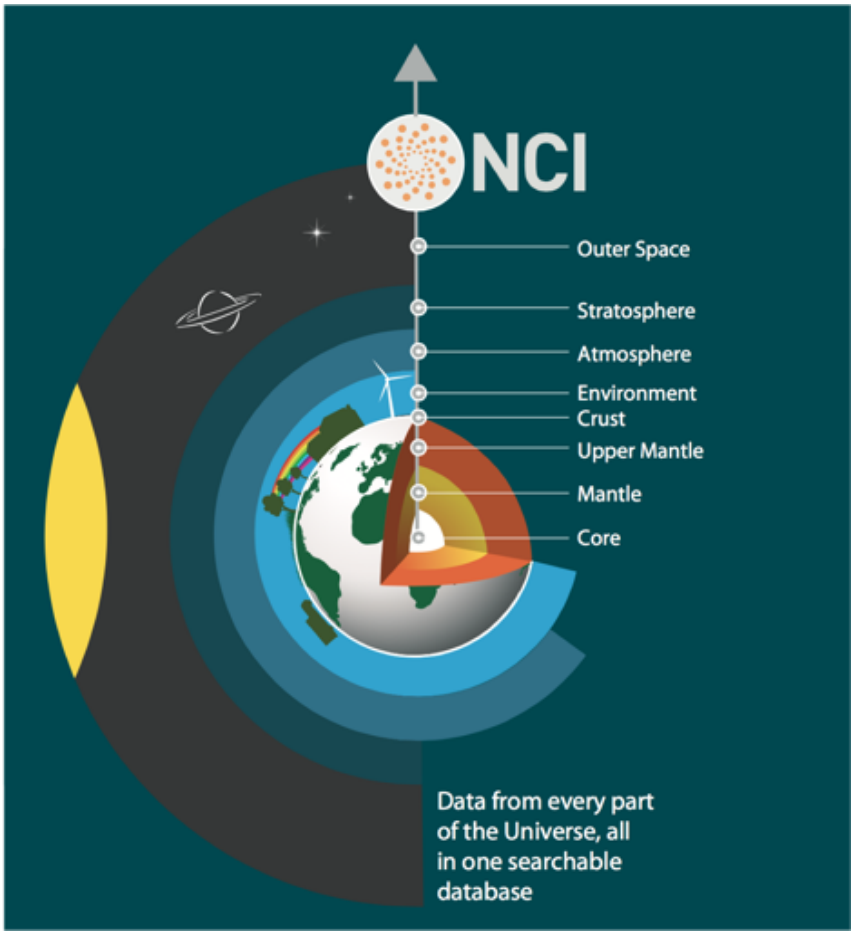
chunked

Tuning parameters affecting performance:

- Order of dimensions
  - last dim is contiguous, rest are stridded.
- Chunk shape
  - Dramatically speeds access along chunked dimensions.
- Compression (per chunk)
  - Reduces size but slows access to individual chunks

Source: Unidata https://www.unidata.ucar.edu

Why important?



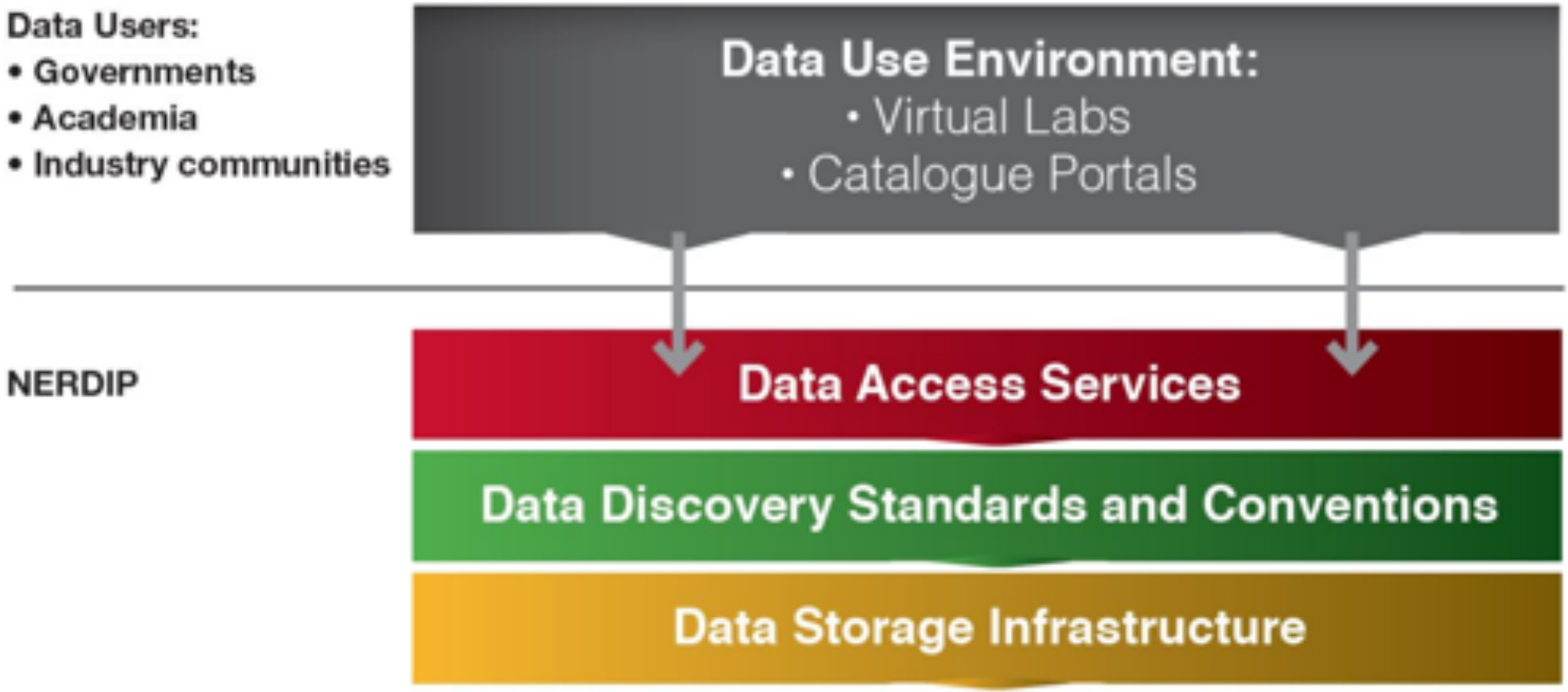
- Application of community-agreed data standards to the broad set of Earth systems and environmental data that are being used

- Within these disciplines, data span a wide range of:
  - Gridded
  - Non-gridded (i.e., trajectories/profiles, point data)
  - Coordinate reference projections
  - Resolutions

Very quick overview of how Data Standards become an important aspect of data:

- Usability
- Interoperability

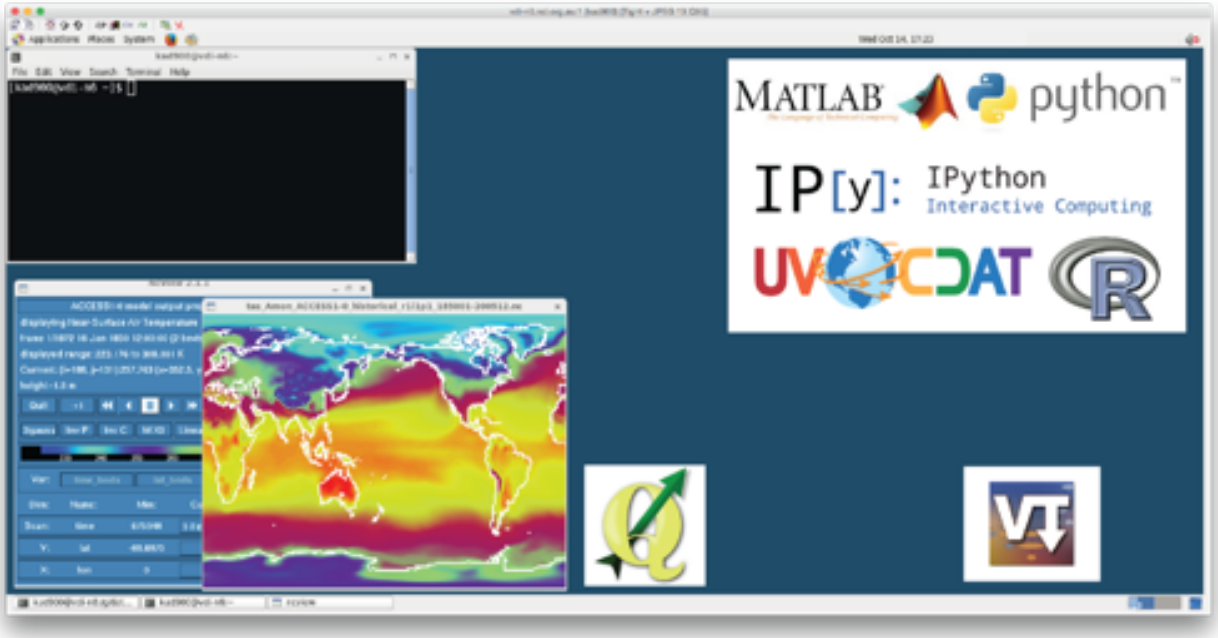
National Environmental Research Data Interoperability Platform (NERDIP)

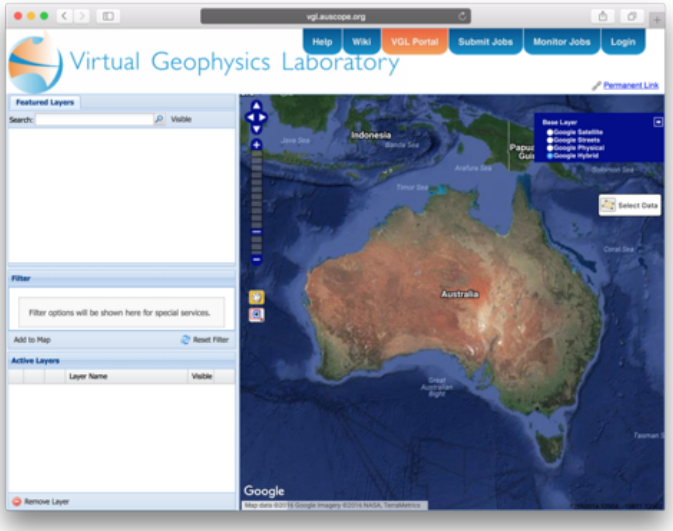


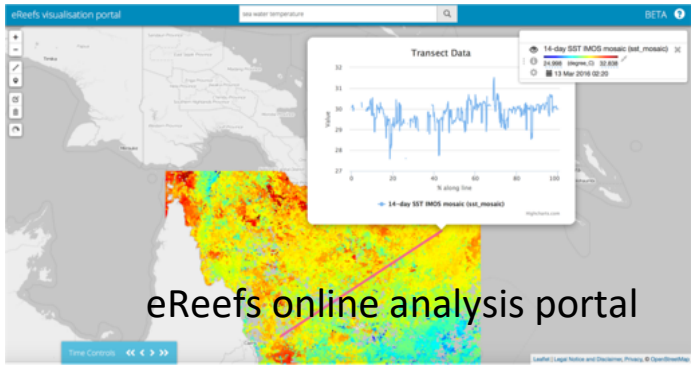
Enables broad range of access

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)







eReefs online analysis portal

New HPC machine is coming!

- ***NCI will NOT migrate the contents of Raijin /short to the new HPC system. All users are strongly encouraged to start archiving and cleaning up files in their /short directories as soon as possible. The Raijin /short file system is provided for temporary files only and is not backed up.***

- ***All NCI users will be asked to update their account passwords and acknowledge current terms and conditions of access for NCI before they can access the new system.***

- ***Recertification is important for account and data security. It ensures only approved users have access to NCI systems and data, and that NCI and its users meet our mutual obligations under Commonwealth legislation, such as the Autonomous Sanctions Act and Defence Trade Controls Act.***