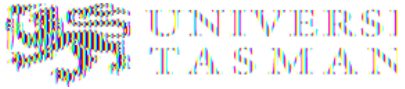
Intro to NCI Data Collections

Data Science Week 2021

NCI AUSTRALIA

Special thanks to all supporters:



Key Takeaways:

1. How to search and browse the NCI Data Catalogue
2. How to access the reference collections and datasets at NCI

NCI Data Collections

- Host Australia's largest research collection of climate, weather, Earth systems, environmental, satellite, and geophysics research datasets.
- Data is a mix of nationally generated datasets as well as replicated international datasets that need to be hosted at NCI for local analysis.
- There are currently more than 13 PB of nationally and internationally significant datasets, and growing…
- As well as data being available more generally, one of the important aspects about this data is that it is organised next to high performance computing and data analysis systems.

NCI AUSTRALIA

NCI Data Collections

NCI supports a number of key internationally recognised data principles:

- **FAIR** data principles for its major data collections. FAIR is Findable, Accessible, Interoperable, Reusable.
- Programmable and high performance access
- Open as possible, Closed as necessary
- Use Data Standards wherever possible
- Transdisciplinary access

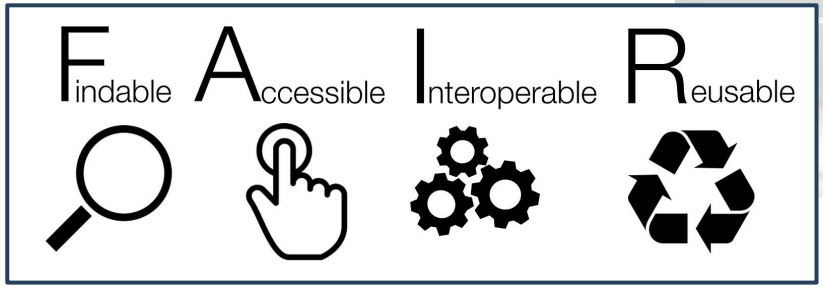
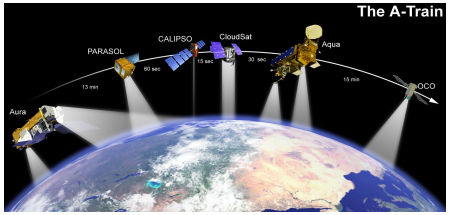
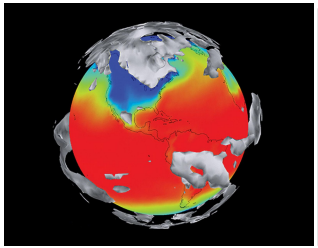


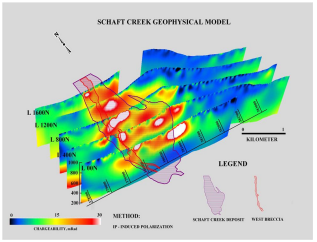
Image credit: SangyaPundir

More background can be found on our wiki: https://opus.nci.org.au/x/NAOVAg

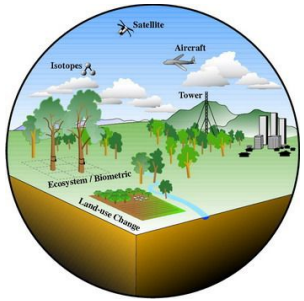
NCI makes available national reference datasets – especially those produced by the government agencies. It is brought together at NCI and organised for both high performance computation & high performance data analysis, as well as making available more broadly to the research community.







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



Browsing and searching for data at NCI

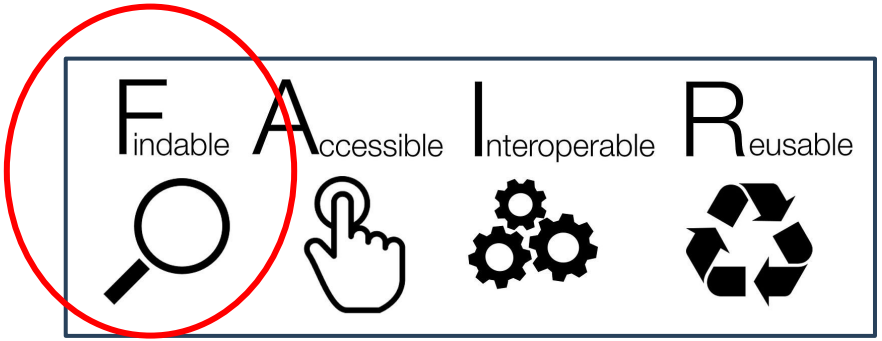


Image credit: SangyaPundir

NCI Data Catalogue

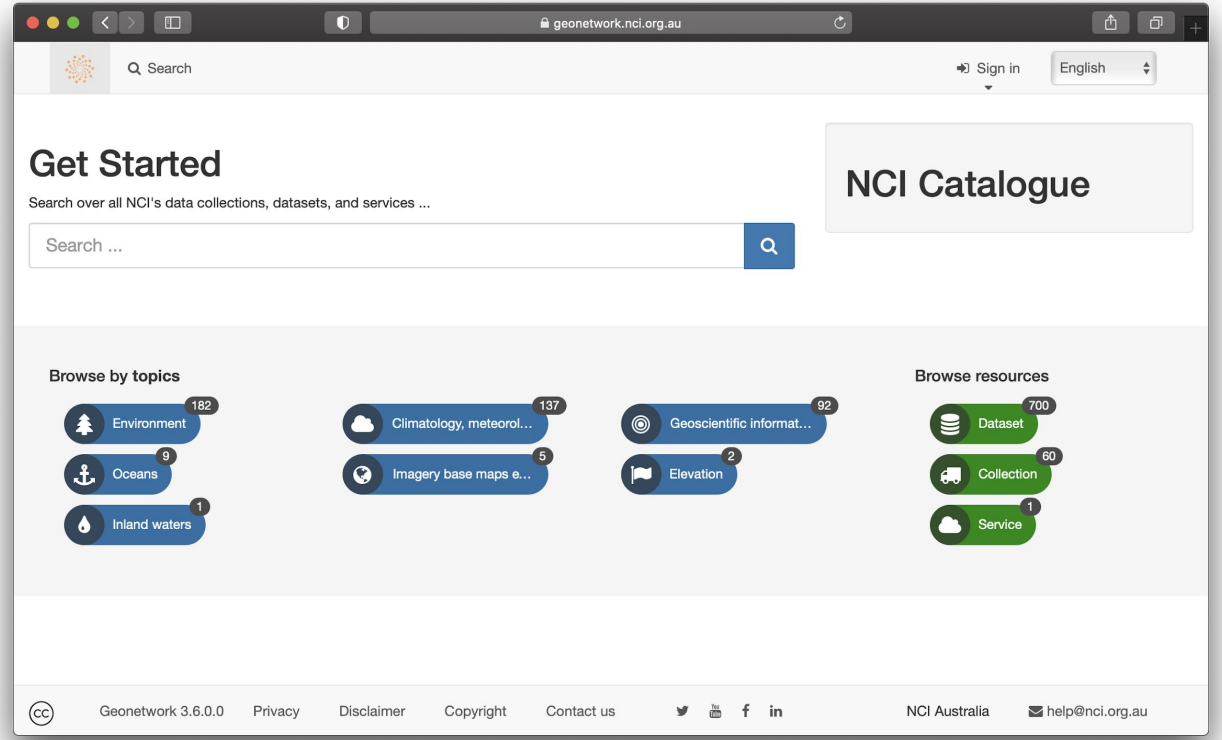
While the Data Catalogue is open and many datasets are available through public services, a number of data collections are only accessible on the NCI systems (i.e., Gadi, VDI) or are more useful within the NCI computational systems. This requires you to complete the NCI new user registration prior to requesting access to a data collection project.

Data Catalogue User Guide

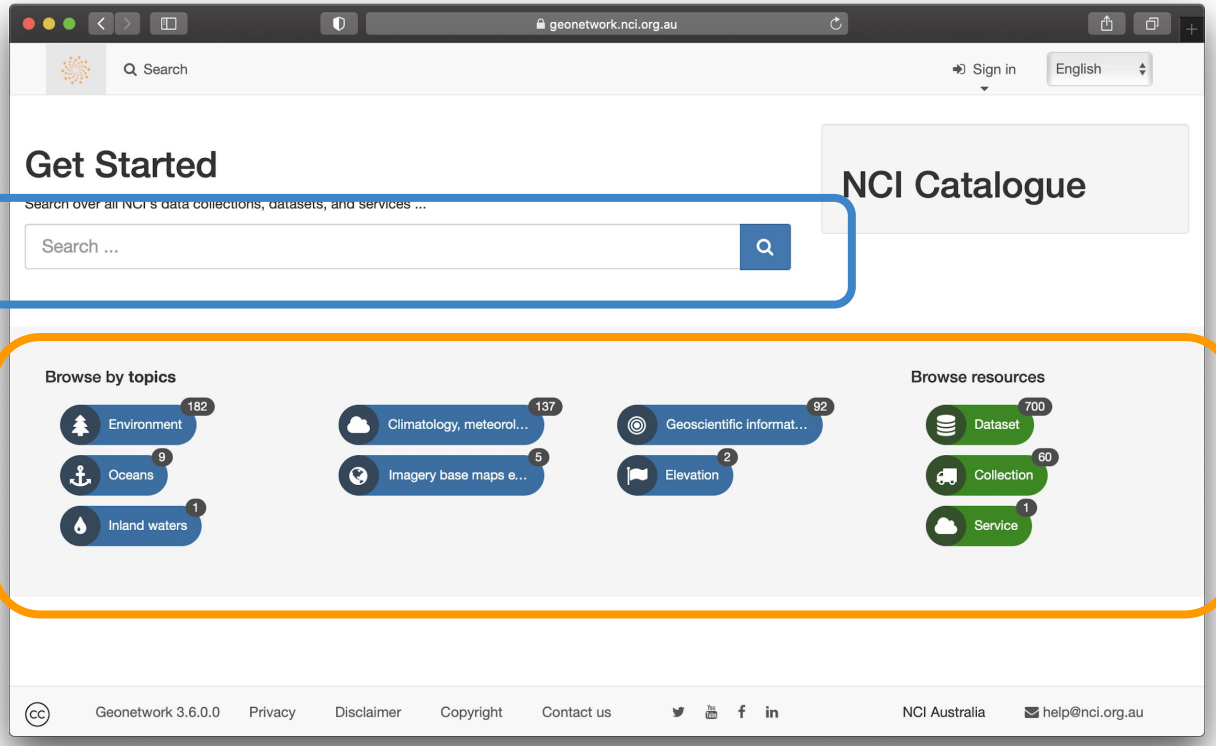
NCI Data Catalogue

You can find information about the NCI data holdings in our Data Catalogue:

https://geonetwork.nci.org.au



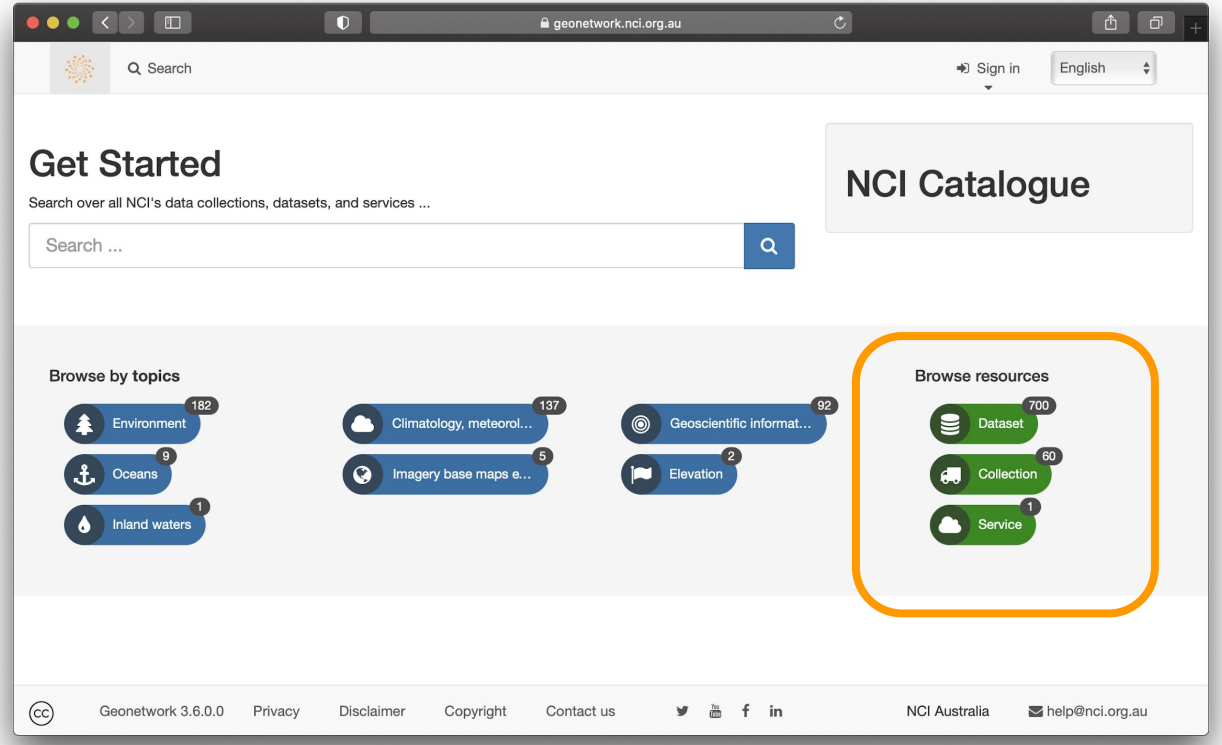
Browsing and searching datasets



Search

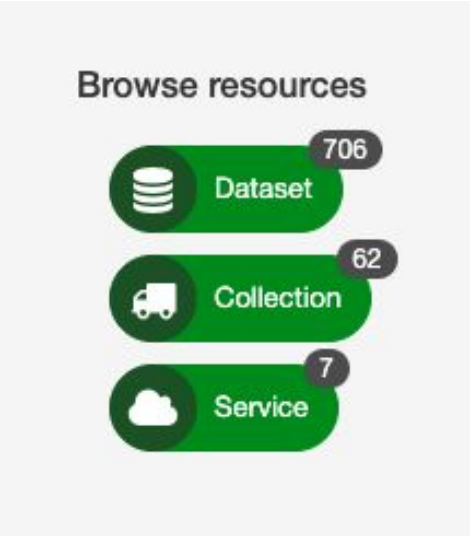
Browse

Browsing and searching datasets



Browsing and searching datasets

Several types of resources you can browse by



NCI Data Collections

- A ***data collection*** is the highest in the hierarchy of data groupings at NCI. It is comprised of either an exclusive grouping of data subcollections; or, it is a tiered structure with an exclusive grouping of lower tiered data collections, where the lowest tier data collection will only contain data subcollections.
- A ***data subcollection*** is an exclusive grouping of datasets (i.e., belonging to only one subcollection) where the constituent datasets are tightly managed. It must have responsibilities within one organization with responsibility for the underlying management of its constituent datasets. A data subcollection constitutes a strong connection between the component datasets, and is organized coherently around a single scientific element (e.g., model, instrument). A subcollection must have compatible licenses such that constituent datasets do not need different access arrangements.
- A ***dataset*** is a compilation of data that constitutes a programmable data unit that has been collected and organized using a self-contained process. For this purpose it must have a named data owner, a single license, one set of semantics, ontologies, vocabularies, and has a single data format and internal data convention. A dataset must include its version.
- A ***dataset granule*** is used for some scientific domains that require a finer level of granularity (e.g., in satellite Earth Observation datasets). A granule refers to the smallest aggregation of data that can be independently described, inventoried, and retrieved as defined by NASA. Dataset granules have their own metadata and support values associated with the additional attributes defined by parent datasets.

More background can be found on our wiki: https://opus.nci.org.au/x/NAOVAg

Collections

These records will:

- Tell you the information relevant to all datasets or products within a collection.

- They will also list these products within the record under a section ‘Associated Records’

Collections

Good for:

- High-level browsing, especially for collections with large number of datasets

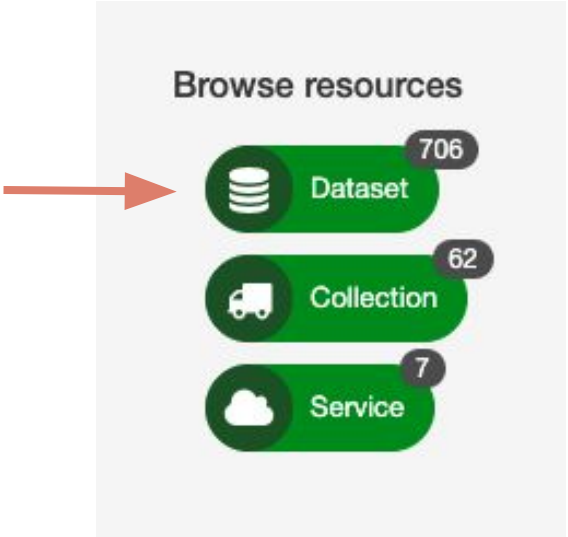
Drawback:

- Dataset specific information will not be displayed on these pages

Datasets

Data records will contain more detailed information about the data product (or sometimes groups of products). This will include metadata items such as:

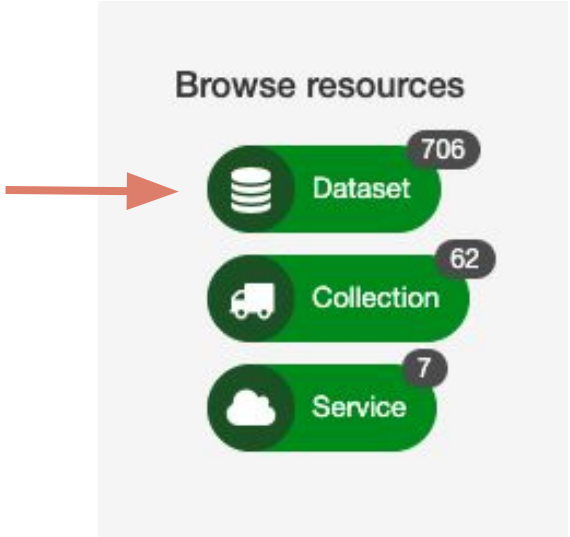
- Abstract
- Data access
- Licensing
- Data lineage
- Data format



Datasets

Good for:

- Dataset specific information such as data access, dois, versioning, data lineage, etc.

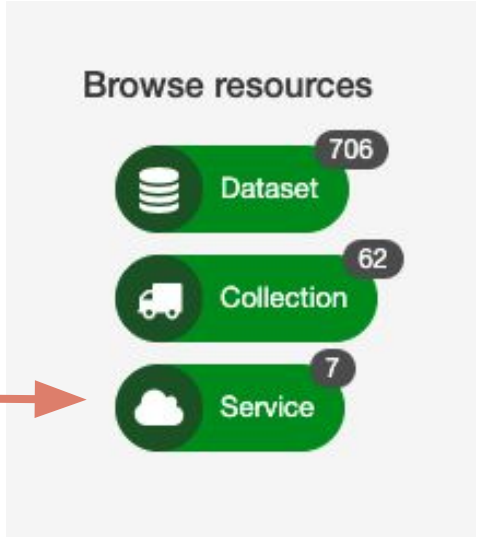


Drawback:

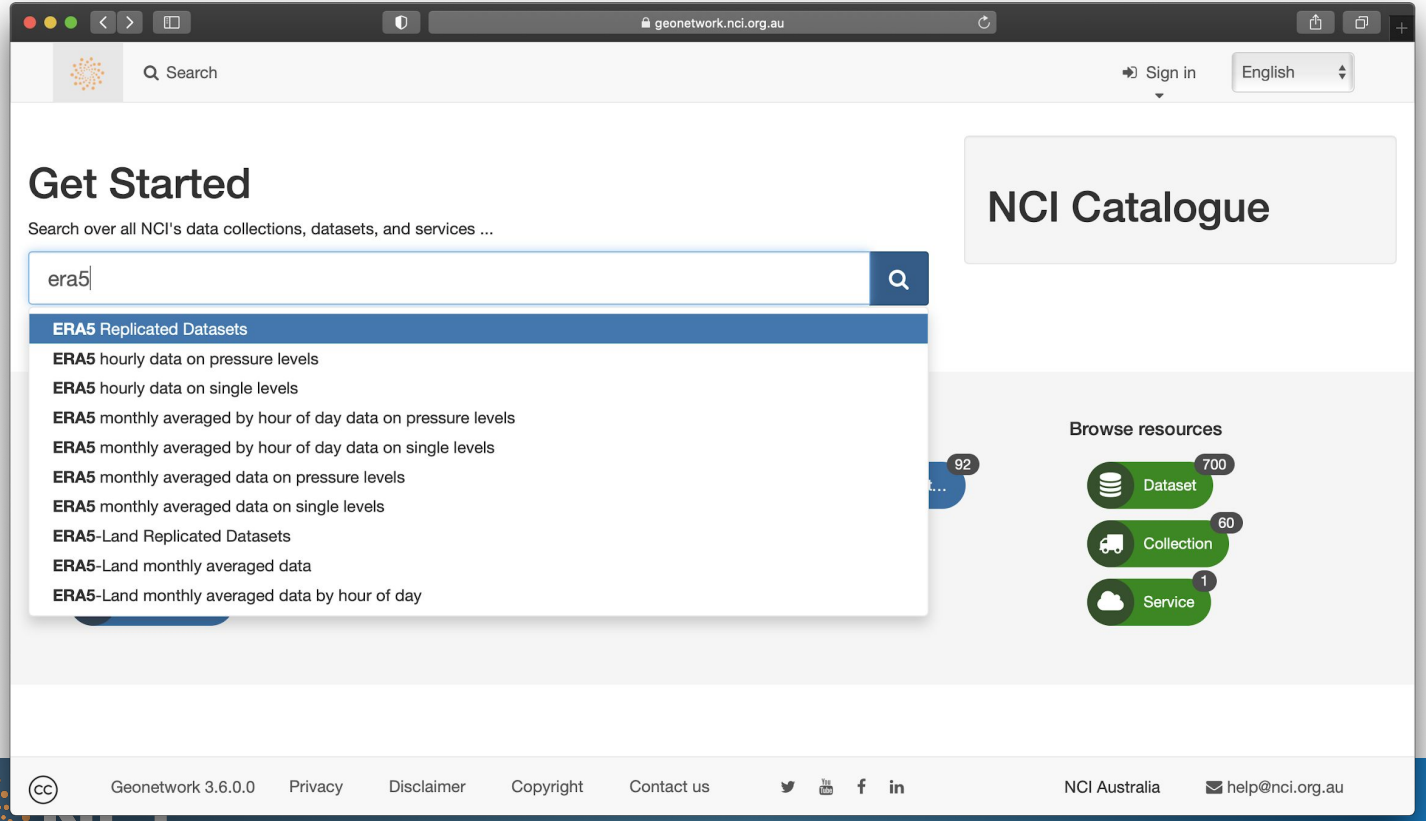
- Lots of them so harder to browse.
- Geonetwork search function has some limitations.

Services

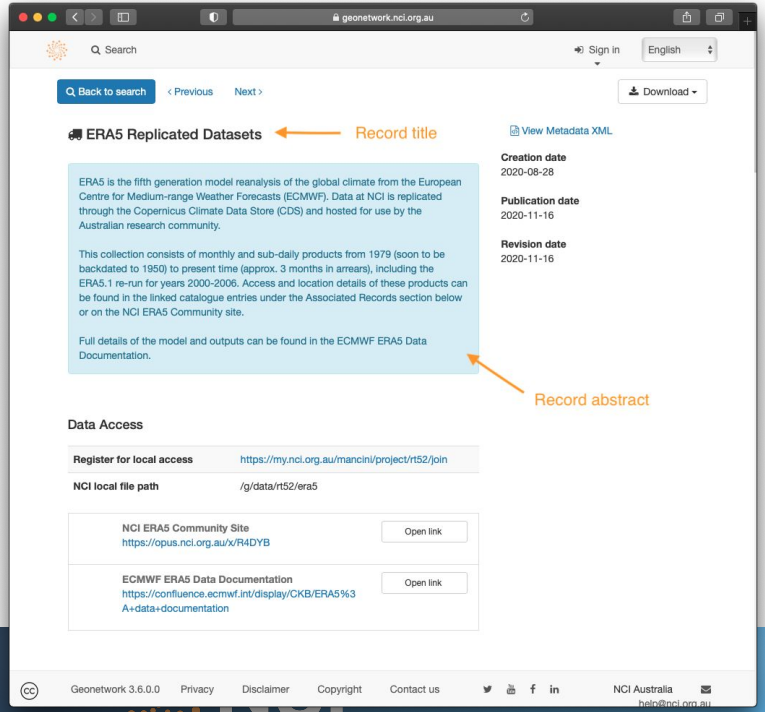
- Service records contain information about a data service provided by NCI.
- These records will link to background information about the service and can be used when citing data access, systems or environments at NCI.



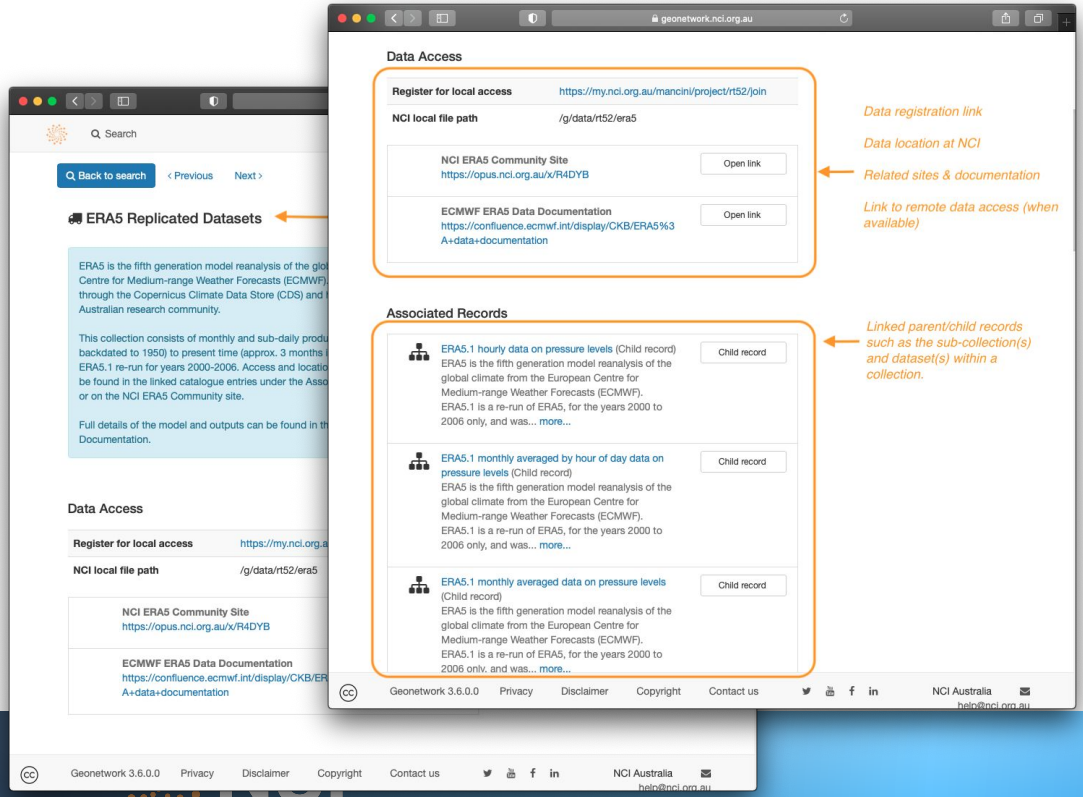
Example 1: Browsing the ERA5 Data Collection



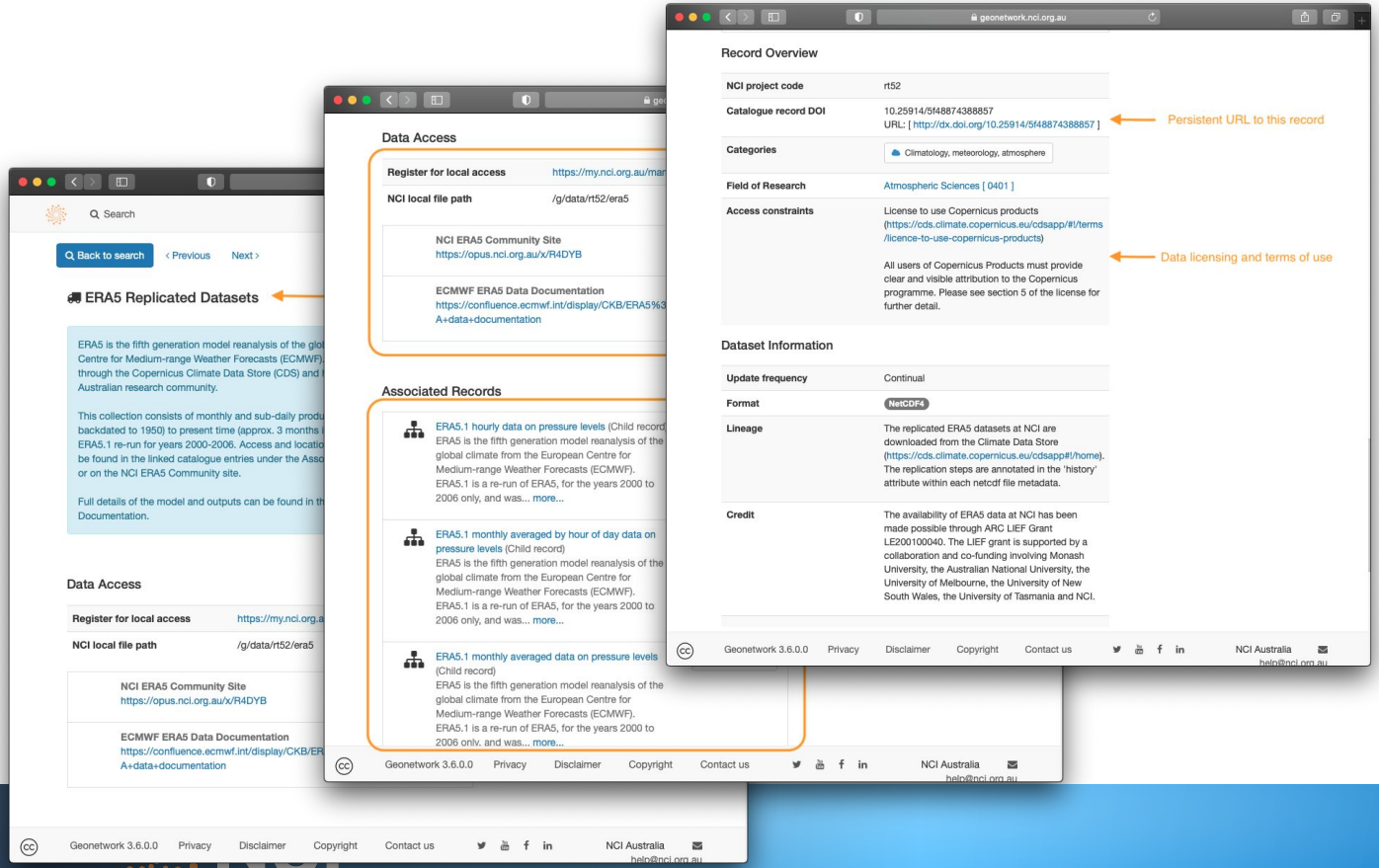
Record metadata



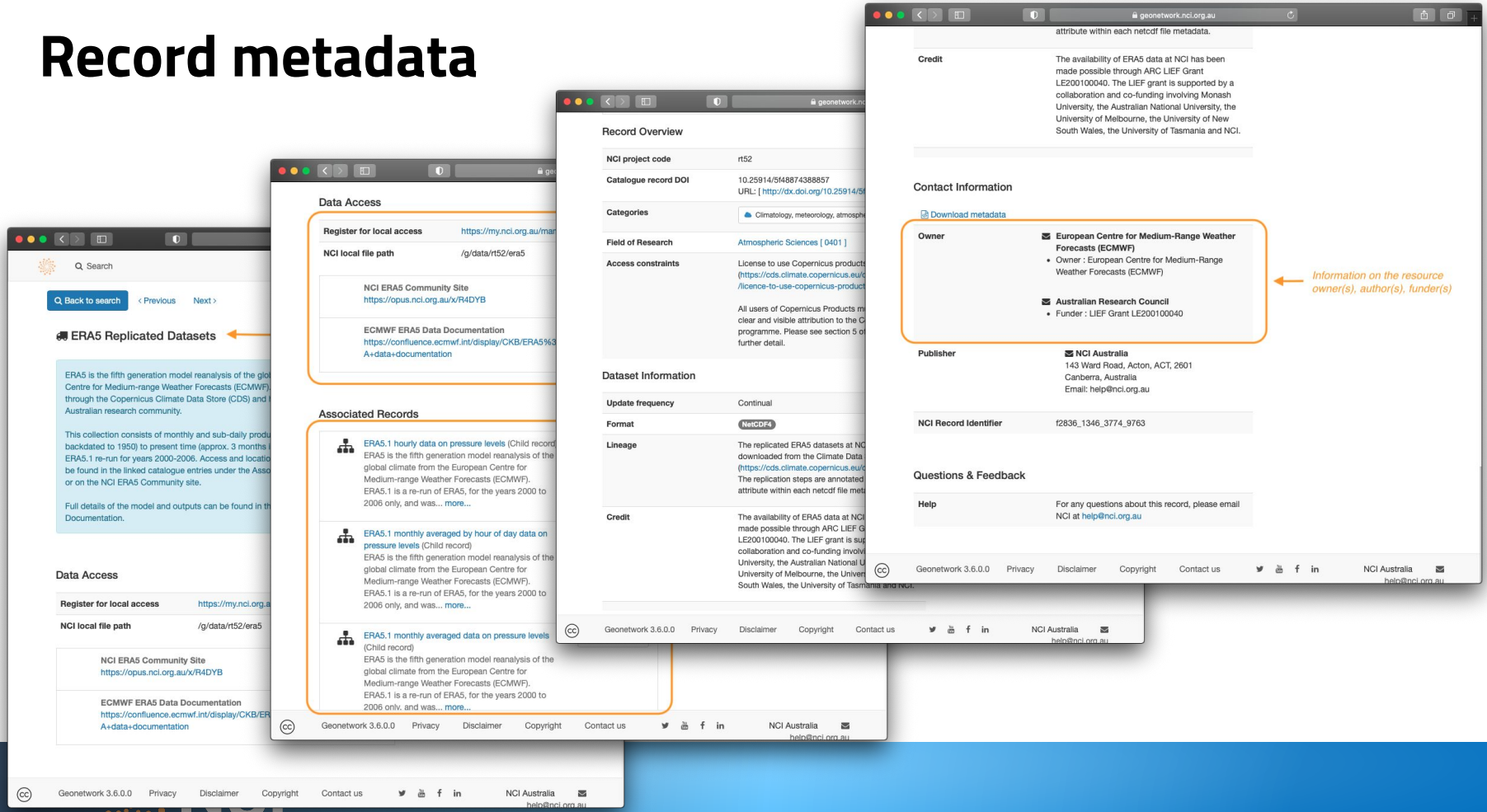
Record metadata



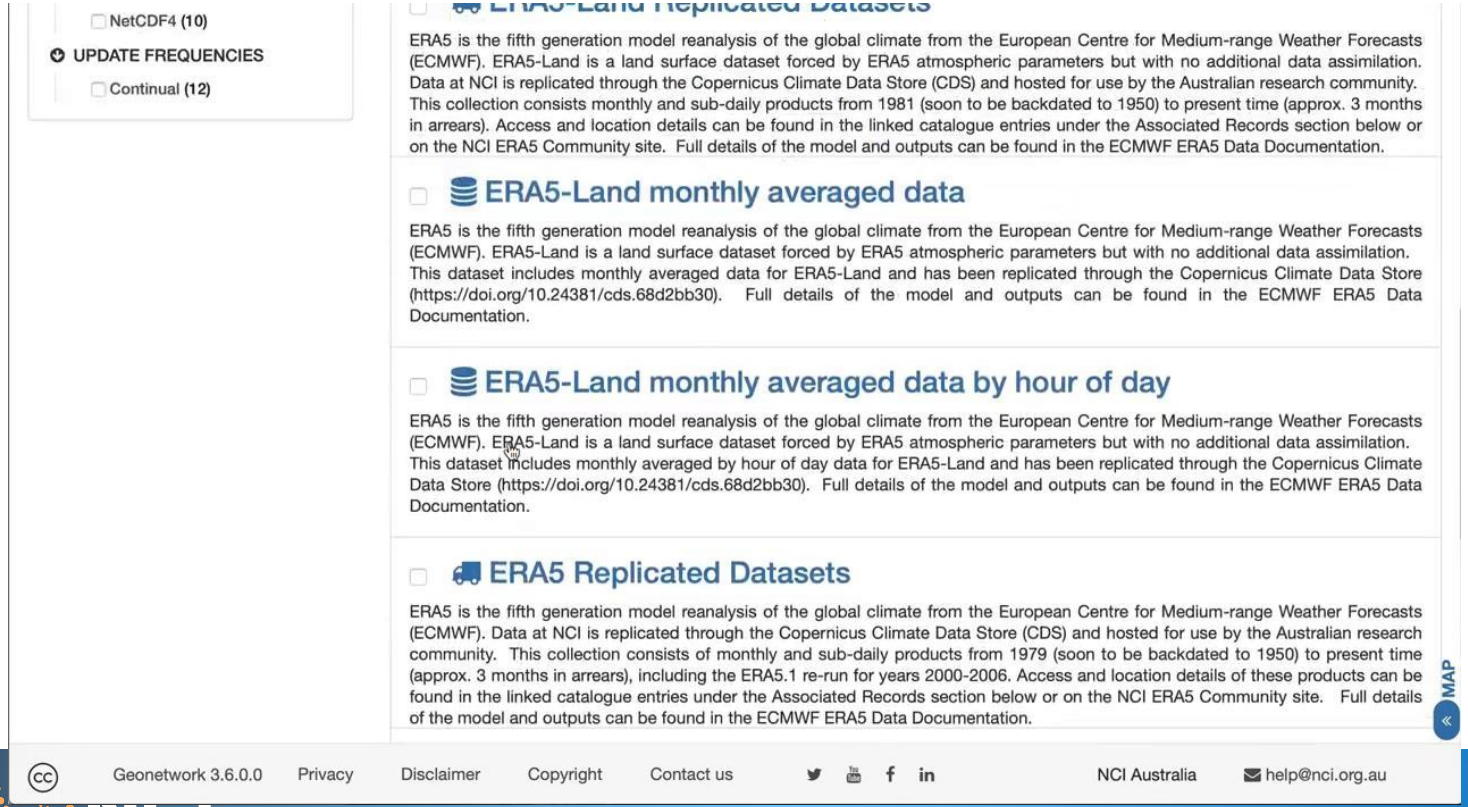
Record metadata



Record metadata



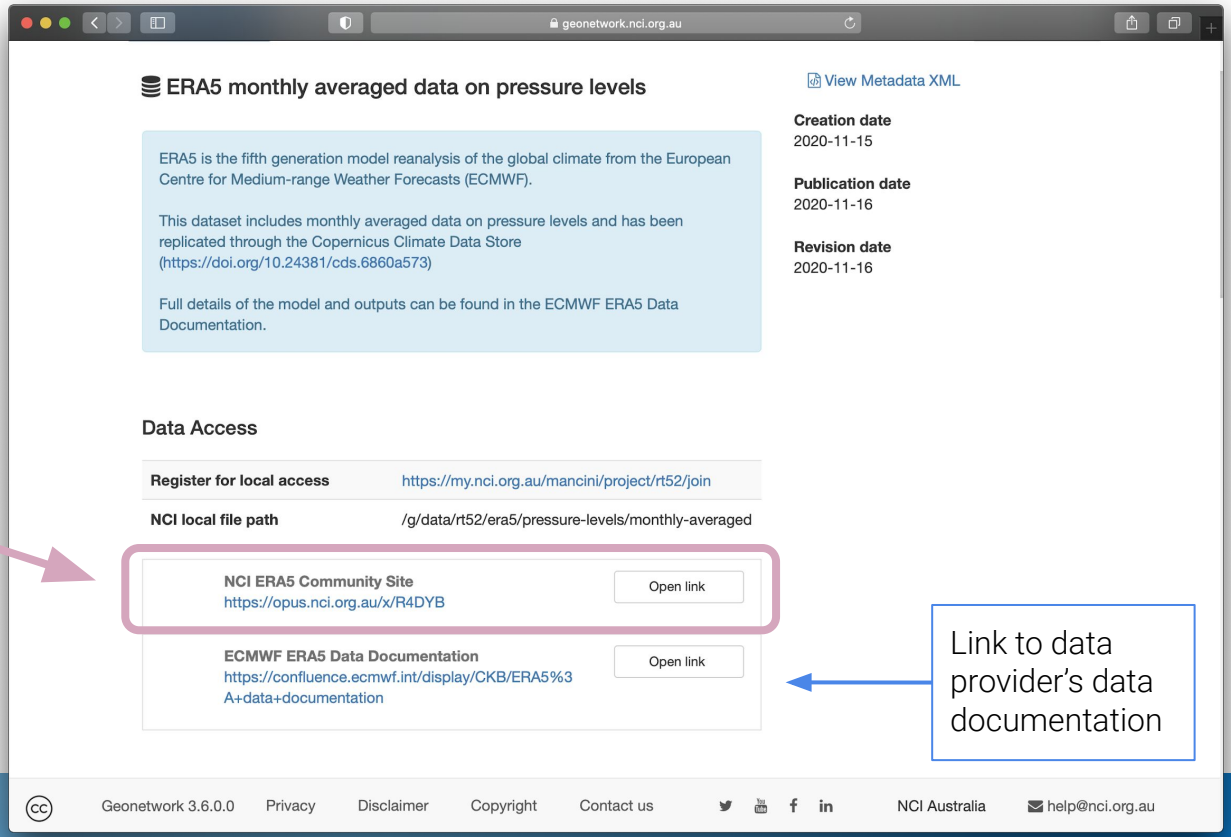
Example: Browsing ERA5



ERA5 Community Site

Further information can be found here, including:

- Data availability (down to specific variables)
- Data organisation
- Replication plan
- Known local issues

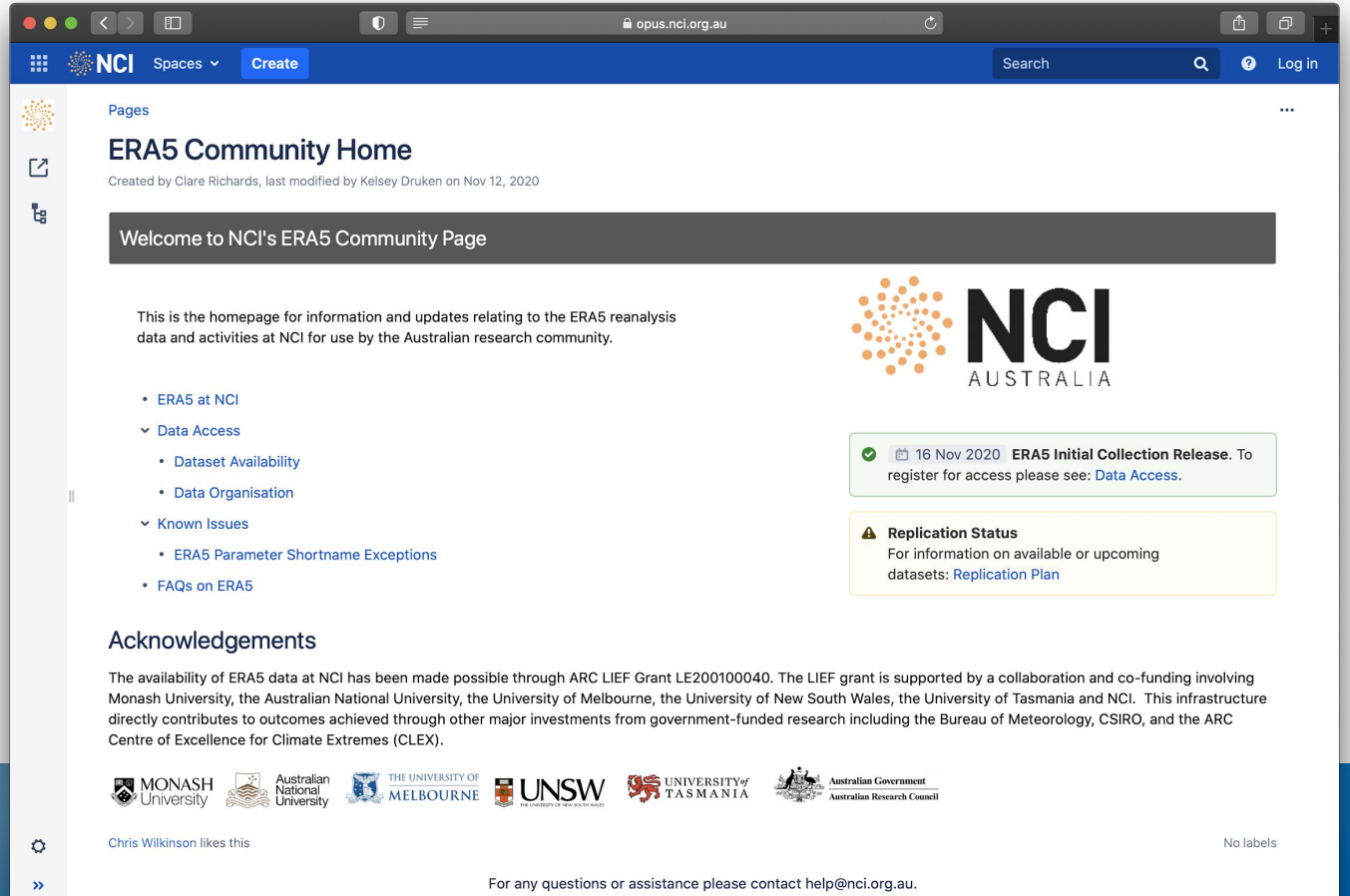


NCI AUSTRALIA

ERA5 Community Site

Further information can be found here, including:

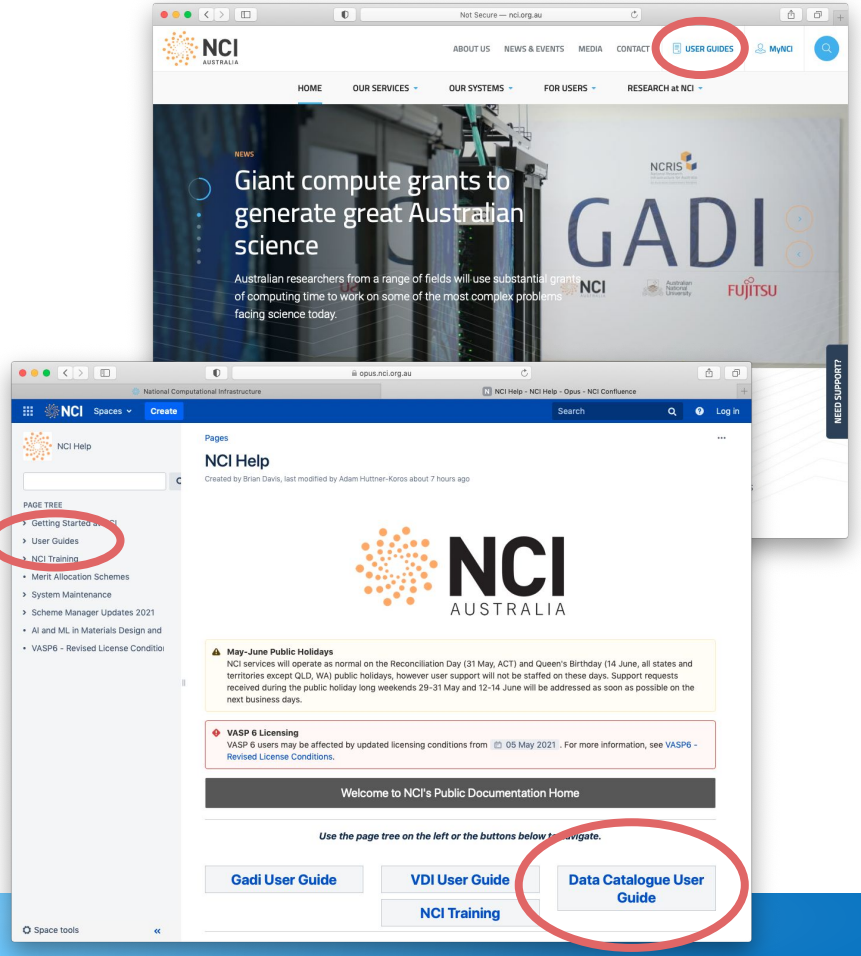
- Data availability (down to specific variables)
- Data organisation
- Replication plan
- Known local issues



NCI AUSTRALIA

Exercise 1

- From the NCI homepage: https://nci.org.au/, select the **User Guides** link in the top menu
- Select the **Data Catalogue User Guide**
- Take a moment to browse some of our collections, datasets and services.



Catalogue homepage: https://geonetwork.nci.org.au/

Accessing data

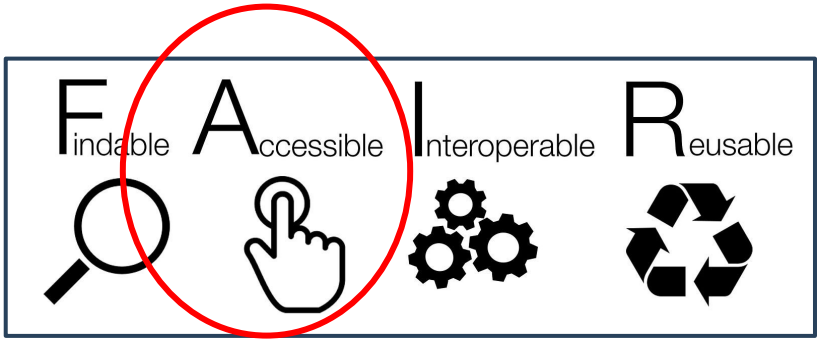


Image credit: SangyaPundir

Accessing data

Several ways you can access the NCI Data Collections:

1. Directly on the NCI systems:
   - Gadi (HPC)
   - VDI (Virtual Desktop)

2. Remotely through data services

Image credit: SangyaPundir

Direct at NCI

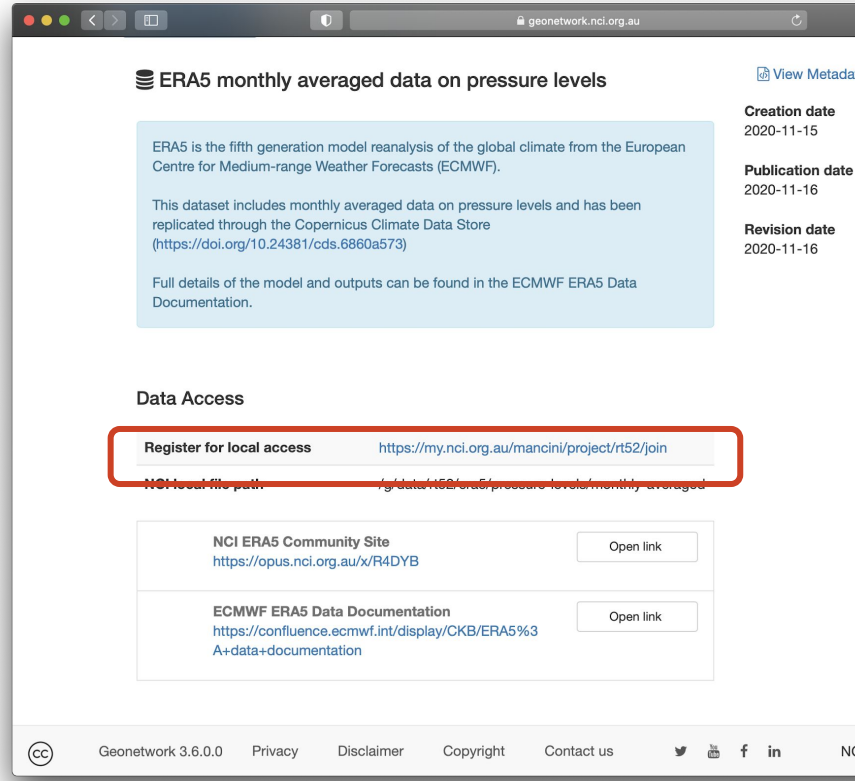
- To access the Data Collections directly at NCI, you'll first need to be a registered user with system resources (this is usually through your organisation or institution).

- For more information on getting started at NCI, please go to our main webpage: https://nci.org.au/users/how-access-nci

- Any questions on how to get started, please reach out to us at help@nci.org.au

Direct at NCI

- Each data collection is hosted within its own project space.
- Users need to register for access to each of these spaces.
  - These registration links are in each of the data catalogue entries.
  - You will be directed to our MyNCI user portal where you will need to agree to the license and terms of use for the collection.

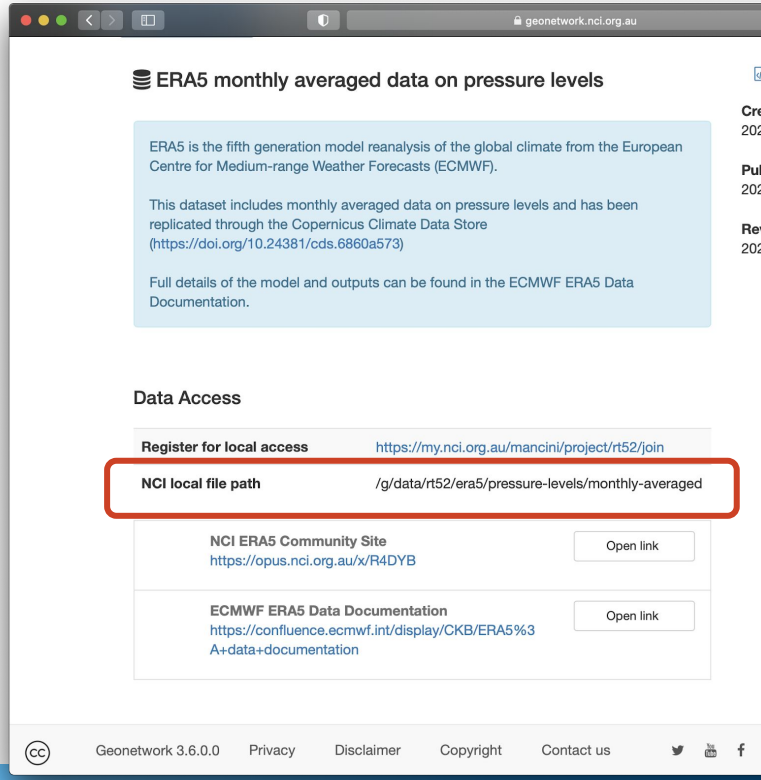


Navigating to the data from Gadi or VDI

Once registered as a member of the project collection code, you will be able to view and access the collection contents from both Gadi and the VDI.

To navigate to the data:

- Locate the collection or dataset location from the catalogue record
- Login to Gadi or the VDI as per usual (please see our User Guides here if you would like more information on getting started to either of these systems)
- Use the 'cd' command to move to the data location:
  - cd <dataset_path>



Accessing data from your Gadi compute job

- When accessing data from your Gadi job you will need to add the data collection code within the '**`-l storage`**' PBS flag.
  - This requirement is new to Gadi.

```
#!/bin/bash


#PBS -l ncpus=48
#PBS -l mem=190GB
#PBS -l jobfs=200GB
#PBS -q normal
#PBS -P a00
#PBS -l walltime=02:00:00
#PBS -l storage=gdata/a00+scratch/a00
#PBS -l wd
```

Data Services at NCI

NCI hosts and supports a wide range of data services:

1. General data servers used for a wide range of collections:
    a. THREDDS - Quick overview here in next few slides
    b. GSKY - Link to user guide can be found in this catalogue record
2. Community specific data services:
    a. NCI ESGF (Earth System Grid Federation) Data Portal
    b. Copernicus SARA Data Portal
    c. Optical Astronomy data services

THREDDS

THREDDS (Thematic Realtime Environmental Distributed Data Services) data server (TDS) developed by Unidata (UCAR) allows for browsing and accessing of data (as well as metadata)

Name	Description
OPeNDAP (DAP2)	Protocol enabling data access and subsetting through the web
NetCDF Subset Service (NCSS)	Web service for subsetting files that can be read by the netCDF java library
Web Map Service (WMS)	OGC web service for requesting static images of data
Web Coverage Service (WCS)	OGC web service for requesting data in some output format
Godiva Data Viewer	Tool for simple visualisation of data
HTTP File Download	Direct downloading

Brief overview of these access methods

Name	Description
OPeNDAP (DAP2)	Protocol enabling data access and subsetting through the
NetCDF Subset Service (NCSS)	ead by the netCDF java library
Web Map Service (WMS)	OGC web service for requesting static images of data
Web Coverage Service (WCS)	OGC web service for requesting data in some output format
Godiva Data Viewer	Tool for simple visualisation of data
HTTP File Download	Direct downloading

Data subsetting options

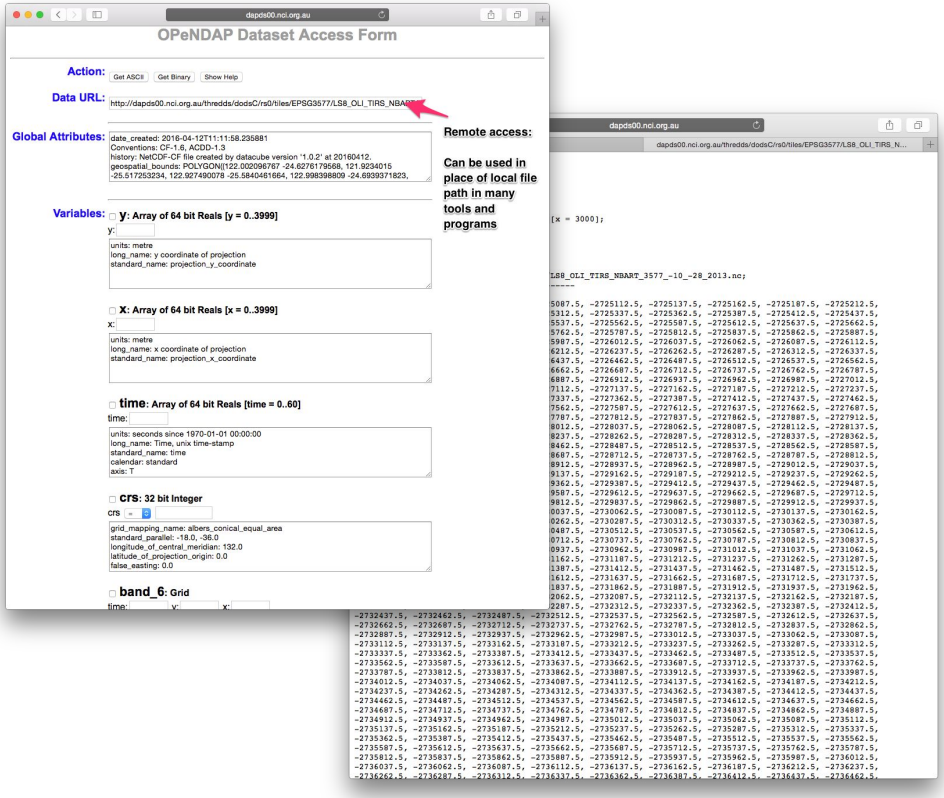
OPeNDAP

Allows for:
Remote access and subsetting. In many tools, the OPeNDAP data URL can also be used in the same manner as in-situ files.

Ideal for:
Remote use with tools and software when working with small subsets of large datasets.

When there is no pressing need to download data.

Possible cons:
Everything is index based.

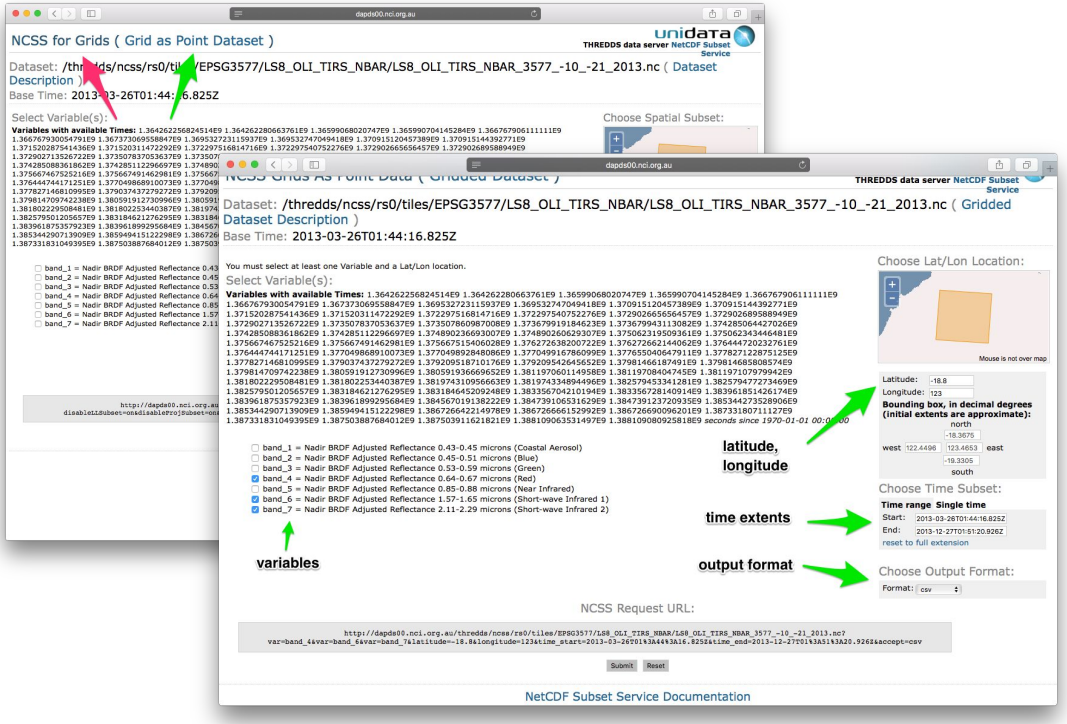


NetCDF Subset Service (NCSS)

Allows for:
Data subsetting of large files based on spatial and/or temporal queries.

Ideal when:
Needing to extract data for specific locations or time. Unlike OPeNDAP, which is all index based, NCSS will do the work to convert the spatial/temporal query into index locations.

Possible cons:
Downloading data locally (when using through the THREDDS interface). Not available on all collections.



New releases

- All new releases are included in the monthly NCI Newsletters as well as in our Data Catalogue User Guide.

- Major release announcements are sent to registered NCI users.

NCI regularly publishes new and updated data collections. This list provides a summary of recent upda
data catalogue at our data catalogue.

April 2021

- COSIMA ACCESS-OM2 0.1 degree global model output (interannual forcing simulation)
- ACCESS NWP APS3 Operational Reference Data Collection
- ACCESS NWP APS2 Operational Reference Data Collection (now available on NCI THREDDS Dat
- Derived Optimal Linear Combination Evapotranspiration - DOLCE v3.0
- EC-Earth3 SSP585 atmospheric forcing dataset for the Ice Algae Model Intercomparison Project experiment v1.0
- Bluelink Ocean Reanalysis - BRAN2020
- ESGF CMIP6 (new datasets added)
- ERA5 (new datasets added)

March 2021

- Seasonal Prediction ACCESS-S2 Hindcast (1981-2018) and Supporting Data Assimilation and Ini
- High-Resolution Modelling of Extreme Storms over the East Coast of Australia v1.0
- Himawari-8 GeoCat 1.0.3 Australian Domain
- ERA5 Replicated Datasets (new products added)
- ESGF CMIP5 and CMIP6 (new datasets added)
- Geoscience Australia Geophysics Reference Data Collection (new datasets added)

Q&A