NCI Data Collections and Publishing

Welcome to NCI's Data Collections information space, where you can find out more about the National Reference Data Collections managed at NCI, the data services we offer, and information about the data publishing process at NCI.

National Reference Datasets

The NCI National Research Data Collection is Australia's largest collection of research climate, weather, Earth systems, environmental, satellite, and geophysics research datasets. NCI also has many other specialised domain datasets, such as optical astronomy and genomic data. This data is a mix of nationally generated datasets as well as replicated international datasets that need to be hosted at NCI.

There are currently more than 13 PB of nationally and internationally significant datasets that are managed at NCI, and with ongoing growth in many of these collections. 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.

Most of the datasets are have been prioritised through the NCI collaboration, particularly in the priority science domains and with NCI partners (ANU, Bureau of Meteorology, CSIRO, GeoScience Australia) and ARC Centres of Excellence. The NCI data collections were primarily funded through a program under the NCRIS Research Data Service (RDS) (2013-7) and subsequently maintained with funding from the Australian Reserch Data Cloud (ARDC). In some cases the datasets have been augmented with a mix of NCI, research collaboration, partner, organisational and ARC funding.

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 where-ever possible
- Transdisciplinary access

Finding and Accessing the datasets published by NCI

You can discover the datasets published and available at NCI using our NCI GeoNetwork catalogue, using ISO19115 compliant data records. As well as the general data catalogue, there are specialist domain information such as the Coupled Model Intercomparison Project (CMIP) service, or Australasia Regional Copernicus Hub. Each collection and constituent dataset has information available as catalogue records in through the NCI GeoNetwork. The data can be accessed through:

- NCI Lustre filesystems /g/data[1a,1b,2,..]/<NCI code>, which are available on NCI's Raijin or VDI systems
- NCI THREDDS data service (http://dapds00.nci.org.au), primarily using Open Geospatial Data Services (OGC) and DAP protocols (e.g., subsetting and aggregation)
- GSKY data service (http://gsky.nci.org.au) using OGC data protocols (WMS, WCS and WPS) for very large datasets (e.g., Satellite imagery)
- Earth Systems Grid Federation (http://esgf.nci.org.au) using DAP protocols
- Sentinel Data service (https://copernicus.nci.org.au/sara.client/#/home)

NCI GeoNetwork provides a discoverability and search portal for these datasets. The NCI code listed for each dataset provides the location in the /g/data Lustre filesystem. NCI account holders logging in need to register for access to the data to help us track demand and communicate information about the data to users and stakeholders. The GeoNetwork records also includes a link to THREDDS data service for data that does not require authenticated access to data. We are progressively adding the location of the data services for these dataset records, including GSKY and other services.

NCI uses internationally recognised Digital Object Identifiers (DOI) on datasets, which can be used to reference these datasets in journal publications or for sharing the location of the dataset landing page. Our goal is to ensure that each dataset lists includes a reference to its license to give confidence around the use of the dataset.

NCI tracks usage statistics around all accesses on datasets - via the open data services and the different protocols of access and usage, as well as in-situ access within the NCI computing systems. This provides information for planning and measuring demand for existing datasets, as well as impacts for upgrades and decommissioning of datasets.

Data management

NCI has a team of expert data managers who work with, organise, and curate the datasets for optimal accessibility, analysis and data publication and accessibility.

Our approach for data management falls into a process of making data fit-for-purpose for computation and programmatic access, and in the context of organising data in the context of a variety of funded schemes. We use the following definitions for this data:

- 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.

These definitions have been described in more detail in a peer reviewed paper on our approach to Quality Data Management. NCI mostly focuses on datasets since it is an more tightly defined data product, and uses subcollections and collections to organise for both data management and licensing requirements.

Preparing and organising datasets

To provide a data publication and sharing service, NCI provides a data management team. This team works closely with data depositors to develop their Data Management Plan that will inform how datasets are catalogued, published, data capacity and managed over time on NCI. This also prepares the information for how the datasets are supported and paid for (e.g., through NCRIS, agency or university funding).

NCI's Data Management Tool portal provides the access point for providing this information. NCI adheres to Implementing a Data Quality Strategy to Simplify Access to Data (link to AGU 2016 abstract)and our Data Quality Strategy to Enable FAIR, Programmatic Access across Large, Diverse Data Collections for High Performance Data Analysis (link to *Informatics* - Open Access Journal).

Help with Accessing or managing datasets

If you represent a university, federal or state government science or institution, NCRIS capability that generates, owns or requires access to big data, contact us at help@nci.org.au to find out how we can help you.