Data Publishing

Introduction

NCI hosts and organises over 40 large data collections that cover a wide range of disciplines. Being co-located with both HPC and cloud facilities, the data collections need to be organised in a systematic way to enable fast programmatic access for in situ analysis across multiple domains, as well as made accessible via data services.

The NCI Data Collections Catalogue manages the details of datasets through a uniform application of ISO19115:2003 - an international schema used for describing geographic information and services. Datasets stored at NCI have to be organised within the NCI catalogue, filesystems and data services in harmonised ways in order to make data accessible using a high degree of specificity and in formats suitable for programmatic (automated) access methods. Hence the organisation and information in the catalogue must be complete and synchronised with the filesystem and data services. Such programmatic access is required by:

- NCI core services such as the NCI supercomputer and NCI cloud-based capabilities;
- Co-located community Virtual Laboratories;
- Remotely, through established standards-based protocols that use the NCI Data Services; and
- Increasingly, through international federations.

This requires data to be well-organised and meet uniform professional standards that makes it usable by programs, developers of programs, and end-users alike. Data also needs to be organised in a way so as to harmonise data operations at NCI, which must publish data simultaneously across several different data servers and services, as well as addressing other data repository management processes and requirements.

Datasets at NCI are primarily ingested and subsequently updated by programmatic means. This can be either through network-enabled replication of datasets organised at other data repositories, or through data generation and/or processing at NCI. Therefore, the organisation of the datasets on the filesystems also needs to be predictable for the publishing aspects of the service, as well as being suitable for automated updating without requiring human intervention.

If you are interesting in publishing your data at NCI or have an inquiry, you can email our NCI Help Desk (help@nci.org.au) or Data Repository team (kelsey.druken@anu.edu.au).

Roles and Responsibilities

NCI is responsible for the quality of the data repository and all its functions and internal consistency of all the information. The following Roles and Responsibilities have been established:

- Data Collections are managed by NCI to agreed community and international standards that strongly relate the data to both transdisciplinary use as well as domain specific needs. NCI leads the process of broader consultation through community management as resolved through the NCI Allocation Committee, its Scientific Assessment Panel and Technical Advisory Group;
- NCI is responsible for the organisation and coordinated activities of data within the collections, in concert with Dataset Managers and Organisational staff such as Data Stewards. This includes development of Data Management Plans (DMPs) and ensuring datasets comply with NCI's Data Quality Strategy; and
- To ensure uniformity in the stakeholder communication and management of the service, NCI is responsible for communications about changes to data areas. The content of advice will be developed in consultation with data providers. It is therefore important that any updates within data areas is managed under controlled procedures.

The value of any data at NCI is considered at the Data Collection and SubCollection level, including funding arrangements for the storage allocations for each of the underlying data Subcollections.

Data Quality Strategy

To ensure that data is managed for all these different uses, NCI Data Quality Strategy to Enable FAIR, Programmatic Access across Large, Diverse Data Collections for High Performance Data Analysis (link to *Informatics* - Open Access Journal) and how we are Implementing a Data Quality Strategy to Simplify Access to Data (link to AGU 2016 abstract) and AGU-DQS-2016-Druken_v2-1.pdf (PDF).

This process is to ensure that datasets complies with known standards, can be delivered using data services and their specialised capabilities, are represented correctly through identified dataa portals of need to the community, and can be used programmatically for high performance simulation and data analysis.

Data Hierarchy Definitions

There are several definitions that are fundamental to how the data catalogue and data directories at NCI are organised: Dataset, Data Subcollection, Data Collection, Data Category and Dataset Granules. While there are a variety of definitions for the terms used that available from other sources, we use those listed below primarily because NCI's focus is on programmatic access to data.

Dataset	A Dataset is a compilation of data that constitutes a programmable data unit that has been collected and organised using the one 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.
Data Subcollection	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 organisation with responsibility for the underlying management of its constituent datasets. A Data Subcollection constitutes a strong connection between the component Datasets, and is organised 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.
Data Collection	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.
Dataset Granule	A Dataset Granule is sometimes used for some scientific domains – particularly in Satellite Earth Observation. In this case it refers to the smallest aggregation of data that can be independently described, inventoried, and retrieved (https://earthdata.nasa.gov/user-resources/glossary#ed-glossary-g). Dataset granules have their own metadata and support values associated with the additional attributes defined by parent Datasets.