



NCI's national environmental research data collection: metadata management built on standards and preparing for the semantic web

Jingbo Wang (1), Irina Bastrakova (2), Ben Evans (1), Kashif Gohar (1), Fabiana Santana (1), and Lesley Wyborn (1)

(1) The Australian National University, Canberra ACT 2601 Australia (Jingbo.Wang@anu.edu.au) , (2) Geoscience Australia, Canberra ACT 2601 Australia

National Computational Infrastructure (NCI) manages national environmental research data collections (10+ PB) as part of its specialized high performance data node of the Research Data Storage Infrastructure (RDSI) program. We manage 40+ data collections using NCI's Data Management Plan (DMP), which is compatible with the ISO 19100 metadata standards. We utilize ISO standards to make sure our metadata is transferable and interoperable for sharing and harvesting. The DMP is used along with metadata from the data itself, to create a hierarchy of data collection, dataset and time series catalogues that is then exposed through GeoNetwork for standard discoverability. This hierarchy catalogues are linked using a parent-child relationship. The hierarchical infrastructure of our GeoNetwork catalogues system aims to address both discoverability and in-house administrative use-cases.

At NCI, we are currently improving the metadata interoperability in our catalogue by linking with standardized community vocabulary services. These emerging vocabulary services are being established to help harmonise data from different national and international scientific communities. One such vocabulary service is currently being established by the Australian National Data Services (ANDS).

Data citation is another important aspect of the NCI data infrastructure, which allows tracking of data usage and infrastructure investment, encourage data sharing, and increasing trust in research that is reliant on these data collections. We incorporate the standard vocabularies into the data citation metadata so that the data citation become machine readable and semantically friendly for web-search purpose as well.

By standardizing our metadata structure across our entire data corpus, we are laying the foundation to enable the application of appropriate semantic mechanisms to enhance discovery and analysis of NCI's national environmental research data information. We expect that this will further increase the data discoverability and encourage the data sharing and reuse within the community, increasing the value of the data much further than its current use.