Linking Derivatives from Big Earth and Environmental Reference Datasets to Publications is not easy: perspectives from the NCI Petascale Data Repository

EGU2019-6377 - X1.180

Lesley Wyborn and Ben Evans
National Computational Infrastructure, Australia *Contact: lesley.wyborn@anu.edu.au

The Emergence of Reference Datasets

Increasingly, large data/HPC facilities now manage major reference collections (e.g., satellite Earth observation, geophysics) in response to a growing research demand at larger geographical scales and/or over longer time periods. Data from multiple surveys and/or progressive time series acquisitions are aggregated into co-located reference collections: some are further standardised and structured into High Performance Datasets (HPD) for use in Data-Intensive Science. Reproducibility requires a capability to be able to access the extract used in research, but at-scale, multiple data exacts cannot be stored indefinitely and there are new requirements from publishers to include persistent identifiers and landing pages for citation.

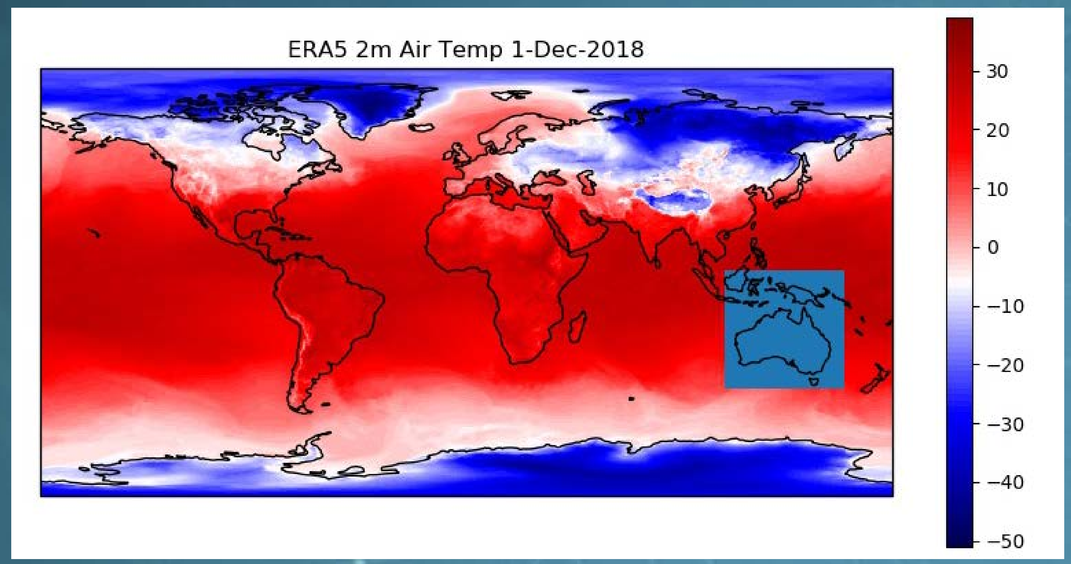
The New Publisher Requirements

The revised Commitment Statement by the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) requires research outputs related to publications to be FAIR and have unique persistent identifiers assigned to each generated/produced dataset. This is easy for small datasets developed by a particular research project over a defined period: outputs are made available for download as files that relate to a publication. Any repository distributing such datasets acts more as a traditional library: with no changes/enhancements made to the dataset.

How Do YOU Cite the Data Derivatives?

Supporting consistent approaches to citation for derivative datasets from large, often dynamic, reference datasets requires community agreement and consideration of costs and technical solutions. Developing such publication standards will not be easy: they have to be consistent across multiple Earth and environmental centres internationally, and include not just research repositories but also government repositories. Increasing dependency by researchers on commercial cloud data storage services such as Google, Amazon, etc., means they might also need to be engaged in this process.

Fig 1. The ERA-5 2m air temperature dataset. ERA-5 provides hourly information at a horizontal resolution of around 31 km globally and at 137 levels in the vertical. The entire dataset from 1950s will be available in late 2019 and is expected to grown to 20 Petabytes. (Graphic from Kate Snow, NCI)



How do YOU preserve your dataset extracts to ensure reproducibility and meet new publisher requirements?

Community discussion is being organised through the Research Data Alliance Data Versioning Working Group here:

WG Homepage
https://bit.ly/2UVJxyd

Please comment on the White Paper being developed here:

WG White Paper
https://bit.ly/2YrZ22O

Perspectives on a Major NCI Collection

Landsat 5 NBAR Pixel Quality 25m v2
Landsat 5 Surface Reflectance NBAR 25m v2
Landsat 5 Surface Reflectance NBAR-T 25m v2
Landsat 7 NBAR Pixel Quality 25m v2
Landsat 7 Surface Reflectance NBAR 25m v2
Landsat 7 Surface Reflectance NBAR-T 25m v2
Landsat 8 NBAR Pixel Quality 25m v2
Landsat 8 Surface Reflectance NBAR 25m v2
Landsat 8 Surface Reflectance NBAR-T 25m v2

Table 1. Multiple derivatives of the Landsat Datasets at NCI

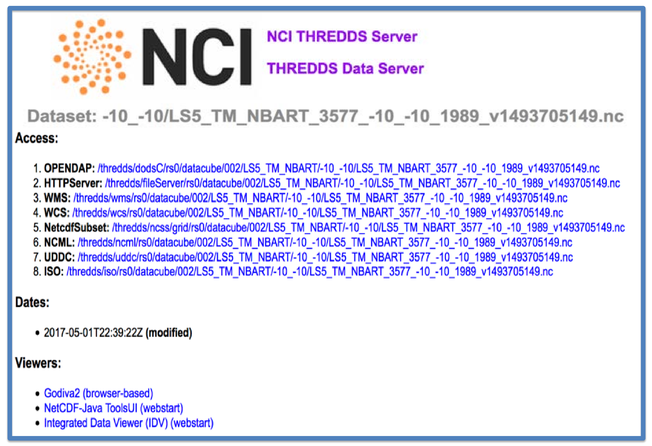


Fig 2. Multiple ways to access and view just one derivative

Potential approaches from NCI

To reproduce a derivative dataset from a national/ global reference dataset requires a dual approach:

1. Researchers need to publish references to all dependent datasets as well as the recipe on how the dataset is derived; and
2. Repositories need to record and make publicly accessible (preferably in machine readable ways) any changes to the dataset as well as referencing the data service(s) that provide data access.

NCRIS National Research Infrastructure for Australia — An Australian Government Initiative
Australian National University
Australian Government Bureau of Meteorology
CSIRO
Australian Government Geoscience Australia
Australian Government Australian Research Council

nci.org.au
@NCINews

Photography encouraged