In the Year of Open-source Science, Which Levels of Data Processing Should We Persist and Make FAIR?

Lesley Wyborn[1], Nigel Rees[1], Rebecca Farrington[2], Arnold Dekker[3]
[1] National Computational Infrastructure ANU, Canberra, Australia, [2] AuScope Ltd, Melbourne, Australia, [3]Sat Dek Pty Ltd, Canberra, Australia.
Contacts: lesley.wyborn@anu.edu.au, nigel.rees@anu.edu.au, rebecca@auscope.org.au, arnoldgdekker@gmail.com

In the year of Open-source Science, researchers are encouraged to make a commitment to the open sharing of any samples, software, data, and knowledge used in any scientific analysis. Most datasets of today begin with raw, full resolution, unprocessed data collected by instruments from either remote (airborne or satellite) or in situ platforms, followed by Analysis Ready Data (ARD), where Raw Field Data (RFD) are calibrated and georeferenced (L0, L1), through to Interpretation Ready Data (IRD), where many data products (L2, L3, L4) can be derived from any L0 and L1 ARD.

Wider availability of co-located data assets and HPC/cloud processing means that the full resolution, less processed forms of observational data can now be more easily processed remotely in realistic timeframes by multiple researchers to their specific processing requirements, which also enables greater exploration of parameter space allowing multiple values for the same inputs to be trialed. The advantage is that better-targeted research products can now be rapidly produced, e.g., using the GADI Supercomputer at the National Computational Infrastructure (NCI) (Table 1).

Processing Level	***No of Sites Processed/CPUs used on Gadi***	***New time to complete work on Gadi***	***Estimated time using parallelised code with 4 cores***	***Estimated time for "site-by-site" processing on average machine with 4 cores***
Packed Raw Time Series Archive	95 sites processed using 96 CPUs	17 minutes	6-7 hours	Multiple days
Level 0 Concatenated Time Series, ASCII	95 sites processed (3328 different days) using 96 CPUs	3 minutes	1-2 hours	Multiple days
Level 0 Concatenated Time Series, netCDF	95 sites processed (3328 different days) using 96 CPUs	2 minutes	1-2 hours	Multiple days to weeks
Level 1 Concatenated Resampled Rotated Time Series, netCDF	Processed 83 sites using 16 CPUs	8 minutes	1-2 hours	Multiple days to weeks

Table 1. Benchmark test for processing different MT time series processing levels using the Magnetotellurics time series data publication (MTtsdp) codes on the Australian NCI Gadi 9.28 Petaflop Supercomputer. The test dataset consisted of MT time series from 95 Earth Data Logger stations with a total of 3328 different days of time series data (Rees et al., 2021, AEGC Conference).

These results show that using HPC and in realistic time frames, many ARDs can be created from each raw field dataset, and likewise numerous IRDs can be created from each ARD. But in Open-source Science, reproducibility and transparency are paramount, and traceability from the raw field data and through to each of the L0-L4 processing levels is critical not only for vouching for the integrity of the derived products, but also to enable attribution and credit to researchers, institutions and funders of the raw field data as well as each level of processing (Figure 1).

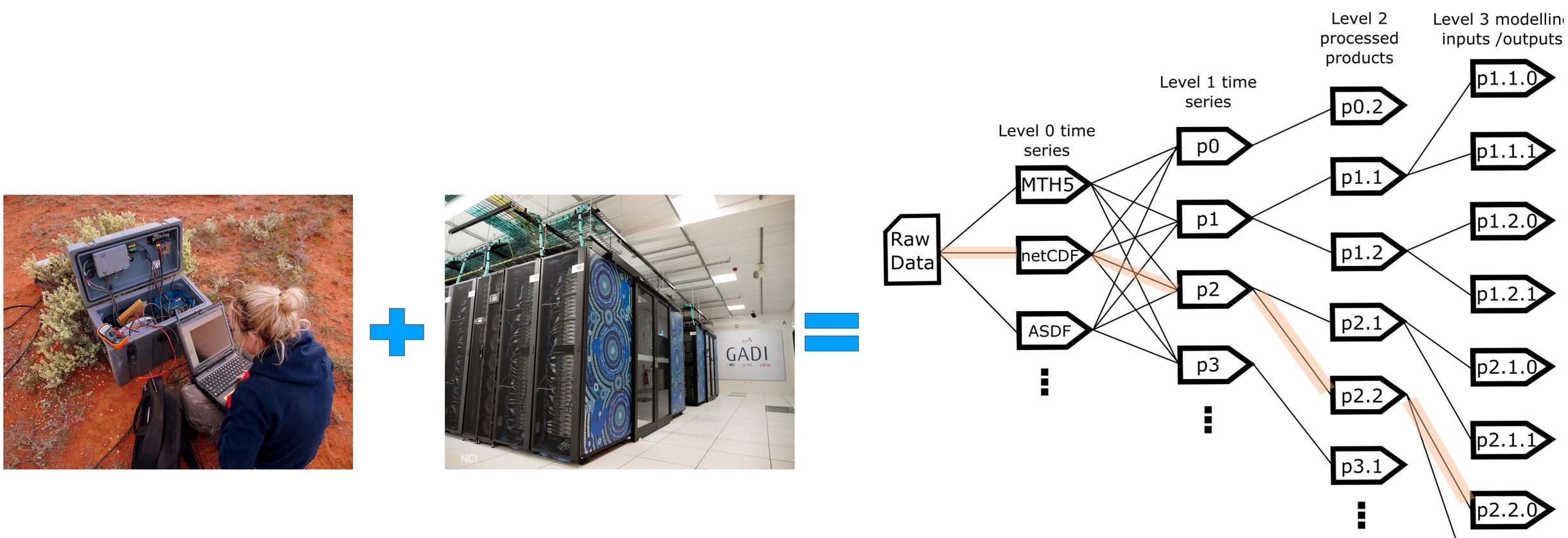

Figure 1. Raw high resolution feld data can be processed using HPC within realisting time frames to generate more accurate products, in multiple formats to suit the user's needs, instead of having to use data products generated by the data supplier that are produced to meet the needs of the majority of users. Through use of identifiers, it is also possible to trace back through each level of processing and the methods that were used at each level of processing.

However, the reality is that in 'Big Data' research areas such as geophysics, remote sensing and climate, where data volumes are measured in TBs and PBs, keeping copies of the raw field data and all subsequent levels of data processing, as well as each individual derivative data product, is not plausible. So how do we decide which raw field datasets and which levels of processing we need to persist and make FAIR? Should we focus on storing and persisting the Raw Field Data and calibrated, minimally processed datasets and then just preserve the workflows and artefacts that created the higher-level Analysis Ready and Interpretation Ready Datasets?

The National High-resolution Geophysics Reference Collections for 2030 project is trialing the use of unique identifiers for each version of each dataset as well as related input artefacts including data, software and tools. It is hoped that the project will lay the foundations for rapid, reproducible, interdisciplinary, in situ analysis of high-resolution national scale geophysical datasets on next generation HPC-cloud systems.

2030 Geophysics Collections Project Web Site: https://ardc.edu.au/project/2030-geophysics-collections/

NCRIS National Research Infrastructure for Australia An Australian Government Initiative
AuScope
ARDC Australian Research Data Commons
tern
NCI AUSTRALIA