AEGC Australasian Exploration Geoscience Conference Brisbane 2021

Using the NCI Gadi Supercomputer to revolutionise processing of MT time series data: results from the GeoDeVL experiment

Nigel Rees[1]
nigel.rees@anu.edu.au

Sheng Wang[2]
sheng.wang@anu.edu.au

Ben Evans[1]
ben.evans@anu.edu.au

Lesley Wyborn[1, 2]
lesley.wyborn@anu.edu.au

Tim Rawling[3]
tim.rawling@auscope.org.au

Bruce Goleby[4]
bruce.goleby@opmconsulting.com.au

Kelsey Druken[1]
kelsey.druken@anu.edu.au

Rui Yang[1]
rui.yang@anu.edu.au

[1]National Computational Infrastructure, Australian National University, Canberra, ACT.
[2]Research School of Earth Sciences, Australian National University, Canberra, ACT.
[3]AuScope, School of Earth Sciences, University of Melbourne, Victoria.
[4]OPM Consulting, Canberra, ACT.

SUMMARY

MagnetoTelluric (MT) time series datasets are expensive to acquire, can be high volume (100s of terabytes), and the time taken to publish (measured from collection to release) often takes more than two years. Time series datasets have been notoriously hard to access: most data providers only make derivative MT transfer functions (EDI files) and model outputs accessible online. Hence, MT practitioners can be reliant on the data processing from raw data to be conducted by others, which may or may not meet their target depth or processing requirements. There is a growing demand for time series datasets to be more accessible to facilitate alternative processing methods, particularly on HPC infrastructures, which enable processing of time series datasets at full resolution and running of larger models with more ensemble members and uncertainty quantification.

To address these issues, the GeoDeVL project experimented with a rapid open, transparent field-to-desktop-to-publication workflow to process and publish MT time series datasets using the new 15 Petaflop Gadi supercomputer at NCI. To do this, parallelised codes were developed to automate the generation of Level 0 to 1 time series data. Creating time series data levels for 95 Earth Data Logger stations now takes minutes, versus days and weeks previously taken using more traditional processing methods.

The process developed under the GeoDeVL project showed how geophysicists can now work with less processed data and transparently develop their own derivative products that are more tuned to the specific parameters of their use case. Further, as new processing methodologies and/or higher capacity computers become available, the rawer forms of earlier surveys are still available for reprocessing. Comparable trials in HPC processing decades ago led to widespread use of HPC in the petroleum exploration industry: will these results lead to similar uptake of HPC in the minerals exploration industry?

Key words: Magnetotellurics, High Performance Computing, data standards, NCI, AuScope.

INTRODUCTION

AuScope is the solid Earth Science capability in the National Collaboration Research Infrastructure Strategy (NCRIS). Its mission is to build an integrated and collaborative research platform that services and facilitates theoretical and applied geoscientific research to investigate Earth Systems and their impact on Australia's liveability, prosperity and environment. In response to the 2016 National Research Infrastructure Roadmap (Australian Government, 2017), AuScope is developing the Downward Looking Telescope (DLT); a distributed observational, characterisation and computational infrastructure providing a capability to image and understand the composition of the Australian Plate with unprecedented fidelity. The AuScope Virtual Research Environment (AVRE) (AuScope, 2020a) is designed to support the DLT as a highly flexible computational environment where researchers can find and access data and tools as online services, and then using notebooks, execute their workflows on a variety of software platforms ranging from personal tablets, through to private/public clouds and world-class High Performance Computing (HPC). The key requirements for data to be consumed by AVRE are that:

1. Relevant solid Earth data from research, government and industry sources are Findable, Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al., 2016) around three integrated networks (geophysics, geochemistry and geology).
2. In alignment with the AuScope 10 year Strategy for 2020-2030 (AuScope, 2020b), data complies with both international data principles, protocols and standards, as well as data management best practice.

The Australian Research Data Commons (ARDC), NCI and AuScope co-funded the 2017-2020 Geoscience Data Enhanced Virtual Laboratory (GeoDeVL) project which was seen as an opportunity to both accelerate the development of the AuScope Geophysics and Geochemistry networks and to foster a change towards a more flexible, virtual research environment that enabled researchers to easily compose their own workflows specifically tailored to their specific research questions. The GeoDeVL project was run as 4 separate, but related work packages: 1) Magnetotellurics (MT); 2) Passive Seismic; 3) Sample Identifiers; and 4) making results accessible in AVRE.

At the start of the GeoDeVL project, MT data providers mainly made processed MT EDI files and model outputs accessible

online as file downloads. The rawer time series datasets were only available through direct request to the author/organisation that collected/owned the data, and often by physical media via the post. A lack of agreed community standards meant that many previously processed EDI datasets from past surveys had inadequate metadata: it was difficult to determine exactly what processing steps had been undertaken to create the EDI files from the source time series files. MT practitioners were thus reliant on the processing conducted by another MT scientist, which may or may not have met their target depth or processing requirements. There is a growing demand for the rawer forms to be more accessible and secure than current practices allow, as they are required not only for replication and reanalysis of published products but also for testing of new processing techniques and computational infrastructures, in particular.

The MT work package of the GeoDeVL project aimed to:

1. Achieve a transparent field-to-desktop-to-publication workflow (Figure 1) that focussed on increasing the transparency and reproducibility of MT data products and enabling linking back to less processed source versions;
2. Investigate how HPC could automate and optimise the processing of the Archived, Level 0 and Level 1 MT time series data (Table 1, Rees et al., 2019);
3. Allow users to undertake more targeted processing of the rawer data forms that was specific to their needs; and
4. Publish new datasets aligned with FAIR principles to increase accessibility and reusability of the data.

However, before the MT time series data could be used in an HPC environment, it needed to be translated into a modern high performance ready, and self-describing format (i.e., netCDF) and made available on both NCI's Gadi supercomputer filesystem and online through web service access (e.g., via OPeNDAP on NCI's THREDDS Data Server). This eliminated the need for local downloading of the large volume time series datasets and enabled users to generate their own transfer functions more tuned to their specific use case; they could also transparently share their workflow via Jupyter notebooks. To meet the FAIR principles, the (meta)data also had to be standardised, particularly with respect to metadata attributes.

This paper focuses on the findings from the MT work package, which was led by NCI, The University of Adelaide, the Geological Survey of South Australia and OPM Consulting, in particular, the substantial transformations of work practices so that the time series data could be accessed by HPC.

METHOD AND RESULTS

The University of Adelaide/AuScope funded Musgraves AusLAMP time series dataset was used to test the field-to-desktop-to-publication prototype on HPC. This dataset consisted of 95 Earth Data Logger stations (3328 different days of time series data) with each station recording one hour blocks of Long Period (LP) electric and magnetic field (EX, EY, BX, BY, BZ) ASCII time series, which meant that there were 120 separate electromagnetic time series ASCII files per recording day. The station metadata for the Musgraves time series were presented in a separate Microsoft Excel spreadsheet.

The data needed to be translated into a self-describing format using agreed metadata standards and formats. Two MT time series standards were available, neither was complete: 1) the Australian MT community draft metadata standard published by Kirby (2019); and 2) the Exchangeable Magnetotelluric Metadata Standard (Peacock and Frassetto, 2020) by the Incorporated Research Institution for Seismology (IRIS) ElectroMagnetic Advisory Committee (EMAC).

The following data profiles of the HDF5 (HDF5, 2021) High Performant Data (HPD) file formats have been used for MT time series data:

1. The Adaptable Seismic Data Format (ASDF) (Krischer et al., 2016) trialled by Geoscience Australia.
2. The Portable Array Seismic Studies of the Continental Lithosphere (PASSCAL) HDF Version 5 (PH5) (Hess et al., 2017) used by IRIS for archiving MT data.
3. MTH5 (Peacock, 2020) used by the US Geological Survey (USGS) for archiving and exchange; and
4. Standard netCDF4 (Rew et al., 2006).

The GeoDeVL project used standard netCDF4, mainly because a, b, and c were still in test/prototype for MT data. The GeoDeVL project automated and optimised the processing and publishing of the archived, Level 0 and Level 1 time series data from the Musgraves Earth Data Loggers instruments. For this, the **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**MTtsdp**) codes (https://github.com/nci/MTtsdp) were developed. The packed raw data and Level 0 MT time series data publishing codes:

1. Produced a zip file of the raw telemetry data for each site logger in the survey (e.g., station1.zip, station2.zip, ...) with no associated metadata;
2. Generated Level 0 concatenated EX, EY, BX, BY, BZ ASCII files at a per station per day granularity. These files do not have any associated metadata available; and
3. Created a single Level 0 concatenated netCDF file per station per day for variables EX, EY, BX, BY, BZ. The MT time series metadata attributes were also added into the netCDF file header.

The Level 1 time series processing pipeline involved:

1. Checking each folder and associated files for any potential issues including incorrect number of samples per day, missing files and evidence of instrument drift;
2. For each station, daily concatenations were merged into a single file encompassing the whole recording interval (minus the first and final day) for each electromagnetic component (EX, EY, BX, BY, BZ);
3. The merged ASCII files were converted into intermediate binary files to accelerate the subsequent I/O operations;
4. Rotation and downsampling (from 10Hz to 1Hz) routines were performed on these binary files based on information provided in the station metadata spreadsheet;
5. For each station, the rotated and downsampled electromagnetic variables and the associated metadata were made into a single netCDF file, i.e., one netCDF file per station over the total recording period.

Once the data were assembled, the processing was undertaken on the 15 Petaflop Gadi supercomputer at NCI which has 155,000 CPU cores, 567 Terabytes of memory and 640 GPUs. Gadi is currently No. 27 on the top 500 Supercomputers list published in November 2020 (Top500, 2020). Considerable performance improvements were achieved as a result of using HPC: processing that previously took days and weeks was reduced to minutes. For example, to generate the Level 0 or Level 1 NetCDF products for 95 sites using 96 CPUs (two nodes: 2x24-core Intel Xeon Platinum 8274 3.2 GHz CPUs per node) on the NCI Gadi Supercomputer took approximately 2 minutes (Table 1).

The archived, Level 0 and Level 1 Musgrave time series datasets were published on the NCI THREDDS Data server (http://dapds00.nci.org.au/thredds/catalogs/my80/AusLAMP/AusLAMP.html), are discoverable in the NCI GeoNetwork Catalogue (http://dx.doi.org/10.25914/5eaa30cc934d0) and are accessible on the NCI g/data filesystem. A user can now easily access and use HPC to process whatever time series data level that suits their needs. Highly parallelised and computationally reproducible workflows from the raw time series right through to the Level 3 modelling outputs could now be developed.

CONCLUSIONS

In this project we have prototyped new and open methods of making MT time series data FAIR and enhanced transparency. This work was to some extent experimental, not just because of the uncertainty of the standards, but also because there were no public domain equivalents to benchmark against. The IRIS consortium is now planning to develop tools for specialised formatting and processing of MT time series data with an intention for these to form the backbone of a long-term, open-source software resource for the MT research community (IRIS, 2020) and are also planning to develop a generalised container for geophysical time series.

The FAIR principles require data to be both human and machine readable, which in turn requires adherence to agreed community standards. There are sufficient standards, crosswalks and protocols for making the data **F**indable and **A**ccessible, but unless there is standardisation on metadata, vocabularies and file formats, **I**nteroperability and **R**euse of MT data is difficult outside of the person/institution that collected/published the data.

One key impact from our field-to-desktop-to-publication prototype is that were an HPC approach to be more widely adopted, the time taken to make MT time series data accessible from collection could be drastically reduced. Currently it often takes more than 2 years for rawer forms of MT data collected with public funding to be made publicly accessible.

The minerals exploration industry has made substantial investments in collecting MT data and hence the key question is whether the HPC developments of the GeoDeVL are transferable to this industry. The MT Musgraves dataset has commonalities with other geophysical time series datasets that sample different physical properties (AEM, seismic, etc). Many of these have the same issue in that the rawer forms of the data are not easily accessible, and the publishing techniques developed here should be transferable to them.

HPC computation is not widely used by the Australian minerals exploration industry. In contrast, the petroleum industry has embraced HPC for processing of geophysics data over the past two decades and at least 12 of the Top 500 supercomputers in November 2020 are focused on oil industry applications. The minerals industry has been slow to adopt HPC techniques, perhaps in part it is because the minerals industry in Australia is dominated by small to medium enterprises that mainly rely on on-premise servers and/or commercial cloud providers with processing and tools provided by third parties. Further, many codes that they use are commercial: few of these are parallelised and optimised for HPC. Hence the preference for many SMEs is for data as more highly processed derivative data products and models: they do not have the capacity to handle the larger volume, rawer time series data. It is hoped that in a successor project, the 2030 Geophysics project (ARDC, 2020), these barriers can be broken down and that together, the research and minerals exploration industry will trial HPC, particularly for effective prospect/district scale targeting. To help mitigate economic, social and environmental risk it is predicted exploration trends may move towards smaller deposits by smaller operators (Moore et al., 2020), which may lead to more dense and voluminous multi-method geophysical surveys aimed at targeting smaller scale anomalies. Access to HPC would enable undertaking of more robust probabilistic analysis on spatially larger and/or higher resolution data sets, which would help build confidence in differentiating anomalies from noise.

RECOMMENDATIONS

We recommend using the EMAC standard (Peacock and Frassetto, 2020) to ensure that Australian MT data collected with public funding can not only be part of global networks, but also that tools and processing software developed anywhere can easily consume the standardised data. Approaches could be made to the instrument companies to make their measurements and outputs compliant with this standard.

The Australian MT community could benefit from greater use of HPC. For national scale surveys such as AusLAMP, the rawer time series levels could be made available in consistent high performant data formats in shared projects on the NCI filesystem, which would enable the MT community to have access to the rawer products on HPC which would help drive more innovative and computational reproducible research.

ACKNOWLEDGMENTS

The GeoDeVL project team would like to acknowledge the assistance, supportive conversations and helpful advice from Kate Robertson, Stephan Thiel (Geological Survey of South Australia); Graham Heinson, Goren Boran, Dennis Conway (The University of Adelaide); Hoël Seillé (CSIRO); Janelle Simpson (Geological Survey of Queensland); Richard Chopping (Geological Survey of Western Australia); Ned Stolz (Geological Survey of NSW); Andy Frassetto, Chad Trabant and Jerry Carter (IRIS, USA); Jared Peacock, Anna Kelbert (USGS); and Kirsten Elger (GFZ, Potsdam).

REFERENCES

ARDC (Australian Research Data Commons), 2020, Combining forces across NCRIS for new insights. Accessed 28 March 2021 https://ardc.edu.au/news/combining-forces-across-ncris-for-new-insights/

AuScope, 2020a, AuScope Virtual Research Environment (AVRE): Data, Visualisation and Analytics. Accessed 28 March 2021. <https://www.auscope.org.au/avre>

AuScope, 2020b, Strategy to address key geoscience challenges. Accessed 28 March 2021. <https://www.auscope.org.au/strategy>

Australian Government, 2017, 2016 National Research Infrastructure Roadmap. Department of Education and Training. <https://docs.education.gov.au/node/43736>

HDF5, 2021, The HDF5 Library and File Format. Accessed 28 March 2021. <https://www.hdfgroup.org/solutions/hdf5/>

Hess, D., Azevedo, S., Falco, N., & Beaudoin, B.C., 2017, PH5: HDF5 Based Format for Integrating and Archiving Seismic Data. AGU Fall Meeting Abstract, 2017, IN42B-03.

IRIS, 2020, IRIS MT Software Development. Accessed 28 March 2021. <https://www.iris.edu/hq/rfp/mtdev>

Kirkby, A., 2019, Developing metadata standards for time series magnetotelluric data. Preview, 2019 (199), 49-53.

Krischer, L., Smith, J., Lei, W., Lefebvre, M., Ruan, Y., de Andrade, E.S., Podhorszki, N., Bozdağ, E. and Tromp, J., 2016, An Adaptable Seismic Data Format, Geophysical Journal International, 207(2). https://dx.doi.org/10.1093/gji/ggw319

Moore, K. R., Whyte, N., Roberts, D., Allwood, J., Leal-Ayala, D. R., Bertrand, G., & Bloodworth, A. J., 2020, The re-direction of small deposit mining: Technological solutions for raw materials supply security in a whole systems context. Resources, Conservation & Recycling: X, 100040. DOI: https://doi.org/10.1016/j.rcrx.2020.100040

Peacock, J., 2020, MTH5. Accessed on 28 March 2021. <https://mth5.readthedocs.io/en/latest/#>

Peacock, J., and Frassetto, A., 2020, A standard for exchangeable magnetotelluric metadata. Accessed March 2021. <https://github.com/kujaku11/MTarchive/blob/tables/docs/mt_metadata_guide.pdf>

Rees, N., Evans, B., Heinson, G., Conway, D., Yang, R., Thiel, S., Robertson, K., Druken, K., Goleby, B., Wang, J. and Wyborn, L., 2019, The Geosciences DeVL Experiment: new information generated from old magnetotelluric data of The University of Adelaide on the NCI HPC Platform. ASEG Extended Abstracts, 2019(1), pp.1-6, DOI: https://doi.org/10.1080/22020586.2019.12073015

Rew, R., Hartnett, E., & Caron, J., 2006, NetCDF-4: Software implementing an enhanced data model for the geosciences. 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography and Hydrology, 6.

Top 500, 2020, Top 500 list - November 2020. Accessed 28 March 2021. <https://www.top500.org/lists/top500/list/2020/11/?page=1>

Wilkinson, M.D., Dumontier M., Aalbersberg I.J. et al., 2016, The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3:160018. DOI: https://doi.org/10.1038/sdata.2016.18

Table 1. Benchmark test for processing different MT time series processing levels using the Magnetotellurics time series data publication (MTtsdp) codes on the NCI Gadi 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.

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

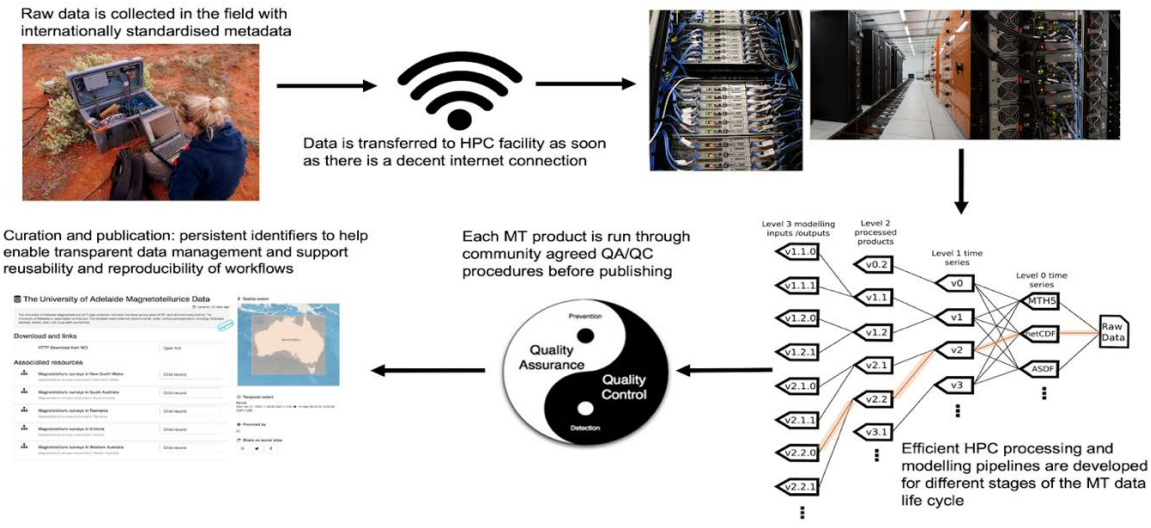

Figure 1: Field-to-desktop-to-publication workflow for magnetotelluric data/metadata.