Building a National High-Resolution Geophysics Reference Collection for 2030 Computation

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

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

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

Rebecca Farrington[2]
rebecca@auscope.org.au

Tim Rawling[2]
tim@auscope.org.au

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

Yue Sun[1]
Yue.s@anu.edu.au

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

SUMMARY

Large volumes of geophysical data have been acquired by universities, industry, federal/state government agencies since the 1950s. However, in many geophysical disciplines the valuable raw time series data has not been made publicly accessible and research geophysicists often have to go through a myriad of hurdles to gain access to raw and time series datasets of interest. In order to increase online collaboration, reduce time for analysis, and enable reproducibility and integrity of scientific discoveries, geophysical datasets will need to evolve to adopt the FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) data principles for both human and machine-to-machine interactions.

The ARDC/AuScope/NCI/TERN funded National High-resolution Geophysics Reference Collections for 2030 Computation Project is working towards making accessible online the rawer, high-resolution versions of AuScope coinvested magnetotelluric and passive seismic datasets, ensuring that they comply with the FAIR data principles and can be integrated with existing government datasets on the NCI HPC platform. Targeted raw geophysical datasets will be ingested and organised at NCI so that they can be (re)processed with computational tools available within the NCI compute systems, and derivative versions of processed data can be linked back to the source datasets. Geophysical data releases will be discoverable through the NCI data catalogue and Research Data Australia: metadata will be structured to enable 'vertical' integration between repositories that have higher level products, but still reference the rawer data at NCI.

The project will make high-resolution geophysical datasets suitable for programmatic access in HPC environments at NCI with the intention of more easily enabling inter-geophysical disciplinary science. Managed geophysical software environments will be created that allow users the ability to fluently scale their Jupyter analysis notebooks to NCI's HPC Gadi system using CPUs and GPUs. This will ultimately lay the foundations for more rapid data processing by 2030 next-generation scalable, data-intensive computation including artificial intelligence, machine learning and data assimilation.

Key words: Geophysics, Time Series, Data Standards, FAIR, High Performance Computing.

INTRODUCTION

Historically, the more voluminous rawer time series forms of geophysical data (processing levels L0-L1 in Table 1 of Rees et al., 2019) which have been acquired by universities, industry, federal/state government agencies, have been hard to access. In part, this is because there is a perception that most users only want calibrated, pre-processed Analysis Ready Datasets (ARD) or visualisations of the datasets in formats that can easily be consumed by online dashboards and GIS products. A more practical barrier to releasing these less processed L0 to L1 time series datasets has been because they are measured in GBs to TBs, whereas higher level products are often merely MBs in volume. Most data suppliers cannot manage and process these larger volume L0-L1 datasets on their current internal infrastructures and the datasets are too large to make accessible as downloads from conventional online catalogues and portals. Although copies of the time series datasets can be obtained by direct request to the data supplier, the average consumer is not likely to have onsite infrastructures suitable for processing. Even on externally available commercial cloud infrastructures, the time series datasets are expensive to store and process, and it is difficult to aggregate multiple datasets to enable processing and modelling at regional to national scales and at high resolutions. As a result, geophysical datasets available online have typically been highly processed and/or downscaled/subsampled products to facilitate online transfer and processing. Further, the majority of these higher level derivative data products rarely have sufficient metadata or provenance documentation and for the user, it was almost impossible to determine what processing steps and input parameters had been used in preparing the released dataset/data product. Thus, practitioners were reliant on the processing of others which may not have been optimal to their target depth and/or their specific processing requirements (Rees et al, in press).

The 2017-2020 Australian Research Data Commons (ARDC), National Computational Infrastructure (NCI), AuScope co-funded Geophysics Data-enhanced Virtual Laboratory (GeoDeVL) Project (Wyborn et al., 2020) focussed on helping accelerate the development of the AuScope National Geophysics Network and fostering a change towards a more flexible, online, virtual research environment, that enabled geophysicists to easily compose their own workflows specifically tailored to their particular research questions. This project prototyped new and open methods of making magnetotelluric (MT) time series data Findable, Accessible, Interoperable and Reusable (FAIR, Wilkinson et al., 2016). Transparent workflows were also developed for processing the rawer, full resolution MT data collected with Earth Data Logger instruments. By adapting MT time series data to modern High Performant Data (HPD) formats and enabling parallelisation of processing codes on HPC, data processing times were reduced from days and weeks to minutes.

As a result of the GeoDeVL, minimally processed L0-L1 geophysical time series datasets are now being progressively organised at the NCI and housed in the AuScope National Research Data Collection. This new, publicly accessible collection is focused on datasets from geophysical surveys whose acquisition, and/or instruments used to collect the data, have been funded through co-investments from AuScope. At NCI, the datasets are collocated with current capabilities for processing and modelling MT data including NCI's High Performance Computing (HPC) infrastructure, HPC/HPD-enabled codes, as well as new capabilities for interactive computing services, Jupyter notebooks and visualisation.

The 2021-2023 ARDC/AuScope/NCI/Terrestrial Ecosystem Research Network (TERN) funded National High-resolution Geophysics Reference Collections for 2030 Computation (2030 Geophysics Collections) Project (ARDC, 2021) aims to both extend the current NCI geophysical data holdings and more importantly, it seeks to position time series geophysical data collections to be capable of taking advantage of the next generation technologies and computational infrastructures of 2030. There are still many unknowns about geophysical processing in 2030, but what is definitely known is that by 2030, exascale HPC resources will be commonplace, data volumes will be measured in Zettabytes (10^{21} bytes) and it will be mandatory for data access to be fully machine-to-machine readable as is also required by the 2016 FAIR principles (Wyborn et al., 2022). Internationally, projects are already exploring geophysical programs that can operate at exascale (e.g., Folch et al., 2022), whilst geophysical research teams from across the globe are beginning to collaborate and develop computationally reproducible narratives on open libraries of multi-physics geophysical data at multiple scales (Figure 1) in-situ on HPC.

This paper will discuss progress on the 2030 Geophysics Collections project during 2021-2022 and will outline the work being done to prepare Data, Software and Computation for these next generation environments.

PREPARING GEOPHYSICAL DATA FOR 2030 ENVIRONMENTS

As a first step towards preparing for 2030 geophysical processing environments, we aim to expand the AuScope National Research Data Collection by publishing raw and calibrated, standardised time series data for magnetotellurics, passive seismic and Distributed Acoustic Sensing (DAS). These datasets will be made accessible online and available for processing on NCI's HPC infrastructure. They will also be interoperable with existing government datasets at the NCI. In order for these national geophysical reference collections to be usable on the exascale computers of 2030, existing datasets will need to be adapted and curated to follow international data standards that are suitable for programmatic access in HPC environments: they will also need to be fully compliant with the FAIR principles and be both human and machine readable. This will require datasets to be accompanied by sufficient machine-readable metadata to enable long-term (re)use and aggregation into multiple, composite datasets that can be progressively updated as new processing methods and tools become available. Processing codes, inversion codes and analysis software will need to be tuned for 2030 compute environments and be carefully managed/versioned to enable users to create transparent workflows from the original time series data through to the transfer functions and modelling outputs.

As noted, for computation at the scales that will be available in 2030, all data will be required to be machine readable. The FAIR principles of Wilkinson et al. (2016) actually explicitly state that data needs to be machine readable, which in turn requires adherence to agreed community formats and metadata standards, preferably those developed and endorsed by international communities. Kelbert (2020) observed that there is a long-standing need to modernise historical MT data formats to a common standard that is fully documented, platform-independent, extensible and accessible to the broader community of geoscientists. Further, many standards and formats currently used in the geophysics community were developed when computational infrastructures were of limited capacity and were dominantly serial, as opposed to the extensive parallelisation that is ubiquitous today. Ip et al., (2019) also noted that due to current format choices, large geophysical data has traditionally been difficult to manage in a consistent, open, and efficient manner. Ip et al. (2019) listed many imitations resulting from these current formats used in Australia including: 1) many working formats are closed and proprietary with potential vendor lock-in; 2) there is limited metadata support and metadata is typically available only as separate "sidecar" files (e.g., .des files for ASEG-GDF); 3) inflexible/opaque binary formats; 4) downloading of the full dataset is typically required and there is no support for subsets via web services; 5) many are ASCII files (e.g., ASEG-GDF) which represents a lowest-common-denominator approach, providing interoperability at the cost of efficiency in both storage and access; and finally, 6) no intrinsic support for machine-to-machine interactions.

For the 2030 project, we are trialling the conversion of The University of Adelaide/AuScope funded AusLAMP Musgraves Province survey time series dataset (https://dx.doi.org/10.25914/5eaa30d63bd17) into the new self-describing MTH5 format (Peacock et al., 2022a) and MT_metadata standards of Peacock et al. (2021). MTH5 is an HDF5 data container for magnetotelluric time series data that uses h5py (Collette et al., 2022) to interact with the HDF5 file and Xarray to link to the time series data. These trials have been tested on the NCI Gadi HPC system which is a highly parallel cluster comprising more than 200,000 processor cores on ten different types of compute nodes.

At the same time, we are attempting to align the legacy AusLAMP Musgraves Province metadata with the international metadata standards for magnetotelluric time series data (Peacock et al., 2021). The new MT_metadata standard is extremely comprehensive, containing attributes that relate to field acquisition through to metadata describing processing of time series data and transfer functions: it is also self describing meaning the metadata are included in the MTH5 file itself and can be parsed programmatically. As noted by Kelbert (2020), historical formats (e.g., Wight, 1988) no longer support the varieties and components of MT data and metadata collected today in modern geophysical surveys and thus, many attributes that were implicitly understood by a small collection of expert MT geophysicists are now explicitly expressed within the new MT_metadata standard.

The AusLAMP Musgraves Province survey consists of 93 Earth Data Logger stations 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. Each station was recording time series data at a frequency of 10 Hz and was left out in the field for anywhere between 5 and 67 days. The station metadata for the Musgraves time series were presented in a separate Microsoft Excel spreadsheet, and each metadata component had to be adapted to comply with the international MT_metadata standard. The ASCII time series were concatenated per run for each station and the total volume of time series data was 106 GB. 90 of the 93 stations were a single run - stations SA246, SA299 and SA324-2 had multiple runs.

We developed code for the creation of:

1. one mth5 file for all 93 Musgraves Province stations
2. one mth5 file per station

Generating one mth5 file for many stations can take a significant amount of time if no parallelism is introduced. For the Musgraves example, if using the mth5 library (Peacock et al., 2022b) alone it would have taken approximately 14 hours to generate our final mth5 file. By utilising h5py's Parallel HDF5 (Collette et al., 2022) and 2 nodes (96 CPUs) on Gadi, we were able to reduce this time to approximately 35 minutes. For the one mth5 file per station model, we were able to create 93 MTH5 files in 3 minutes and 35 seconds using 2 nodes (96 CPUs) on Gadi. Tutorials on both of these workflows were contributed to the mth5 github repository:
https://github.com/kujaku11/mth5/blob/master/docs/examples/notebooks/mth5_in_parallel.ipynb
https://github.com/kujaku11/mth5/blob/master/docs/examples/notebooks/mth5_in_parallel_one_file_per_station.ipynb

The key conclusion from this work on preparing MT datasets for 2030 computation is that the speed at which the time series data can be processed means that the rawer time series forms of the data can now easily be processed in realistic timeframes by individual researchers to their target depth or specific processing requirements (Rees et al, in press). This also enables greater exploration of parameter space and multiple values for the same inputs to be trialled. As multiple products can now be rapidly produced from each time series dataset, each derived product will be uniquely identified using Datacite Digital Object Identifiers (DOIs) and time stamped. This will enable provenance tracking from any raw data acquisition to multiple derivative products and will enable 'vertical' integration between repositories that have higher level products, but still reference the rawer data at NCI. All datasets will be discoverable in the NCI online catalogue (https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/home) and in Research Data Australia (https://researchdata.edu.au/).

COMPUTATIONAL AND SOFTWARE ENVIRONMENTS ON NCI

Geophysical researchers have various options for running their computational workflows at NCI including:

1. Gadi which is Australia's peak research supercomputer, a highly parallel cluster comprising over 200,000 cores including 160 GPU compute nodes that provide 640 NVIDIA V100 GPUs. To run geophysical compute tasks

such as time series processing or inversions on Gadi, users are required to submit jobs to queues and Gadi utilises PBSPro to schedule all submitted jobs.

2. The Australian Research Environment (ARE) which is a web-based graphical interface for performing computational research. ARE provides access to NCI's Gadi supercomputer and multi-petabyte data collections and supports applications such as JupyterLab and Virtual Desktops.

Many modern geophysical workflows utilise open source libraries and tools that are based on the classic core Project Jupyter (Perez & Granger, 2015; Granger and Pérez, 2021) programming languages: Julia, Python and R. To allow for a more fluid experience for geophysical data analysis and processing, the NCI-geophysics module (https://opus.nci.org.au/x/1wG7CQ) has been developed which integrates Python, Julia and R environments together with thousands of pre-built libraries. NCI users can access these libraries directly and use them within their own workflows without having to spend significant time building their own software environments. One advantage of having a shared, quality controlled, community driven, regularly updated geophysical software environment module is that scientific workflows can easily be shared and reproduced without geophysical researchers having to grapple with software engineering principles such as containerisation, package dependencies, package managers (e.g., PIP, Conda, CRAN), compiling for HPC architectures, debugging, PATH variables, etc. With the NCI-geophysics module, it is as simple as running:

```
$ module use /g/data/up99/modulefiles
$ module load NCI-geophys
```

and any user has access to over 1500 Python/Julia/R libraries that can be used either on Gadi or via a JupyterLab or Virtual Environment on the ARE.

The NCI-geophysics module has combined and compiled various open source Python/Julia/R geophysical software packages including ObsPy (Beyreuther et al., 2010), Pyrocko (Heimann et al., 2017), SimPEG (Cockett et al., 2015), MTpy (Krieger and Peacock, 2014; Kirkby et al., 2019), MTH5/MT_metadata (Peacock et al., 2022), HiQGA (Ray et al., 2022), JUDI (Witte et al., 2019), JuliaGeo (Barth et al., 2022) and many others with common data science, data format, HPC scaling and visualisation packages including Dask (Rocklin, 2015), Xarray (Hoyer and Hamman, 2017), Open MPI (Gabriel et al., 2004), mpi4py (Dalcin and Fang, 2021), h5py (Collette et al., 2022), Zarr (Miles et al., 2022), TileDB (Papadopoulos et al., 2016), netCDF4-python (Whitaker et al., 2020), GDAL (Rouault et al., 2022), GeoPandas (Jordahl et al., 2021), Bokeh (Collins et al., 2020), seaborn (Waskom, 2021), Plotly (Plotly Technologies Inc, 2015), JuliaParallel MPI.jl (Byrne et al., 2021), Mackie.jl (Danisch & Krumbiegel, 2021), ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2022), tidyr (Wickham and Girlich, 2022) and many others.

As geophysical datasets are ever increasing in volume and complexity, automating the processing, modelling and interpretations of these datasets with Artificial Intelligence (AI) and Machine Learning (ML) tools is becoming more commonplace (e.g., Puzyrev, 2019; Wrona et al., 2021; Yu and Ma, 2021). To simplify the use of AI/ML tools on HPC, NCI have developed and manage an AI/ML environment (https://opus.nci.org.au/x/c4IYCg) that allows researchers to more easily run their AI/ML workflows using multiple GPU nodes on Gadi. This module has bundled together many of the major data science and machine learning packages including Tensorflow (Abadi et al., 2016), Pytorch (Paszke et al., 2019), Scikit-learn (Pedregosa et al., 2011), Horovod (Sergeev and Del Balso, 2018), Xarray (Hoyer and Hamman, 2017), Dask (Rocklin, 2015) and Ray (Moritz et al., 2018).

In addition to the NCI geophysics and AI/ML modules, standalone open source geophysics related applications that support parallel computation are progressively being added to Gadi including OpenQuake (Pagani et al., 2014), ModEM (Kelbert et al., 2014), esys-escript (Gross et al., 2007; Schaa et al., 2016), Mare2DEM (Key, 2016), Firedrake (Rathgeber et al., 2016), FEMTIC (Usui, 2015) and BIRRP (Chave and Thomson, 2003).

By 2030, the use of well managed and FAIR software environments that integrate a plethora of different complex and intricate open source codes/libraries in multiple languages on HPC will become routine. Many researchers typically just want software environments to "work" on HPC, and do not have the skills or time to manage thousands of software libraries and/or compile/adapt source code for HPC resources. Global and interdisciplinary research teams of the future will become more reliant on software engineers to develop community driven software environments that aid and enhance the transparency and reproducibility of their scientific workflows.

CONCLUSIONS

The ARDC/AuScope/NCI/TERN funded 2030 Geophysics Collection is an exploratory project that is progressively making accessible online a selection of rawer, high-resolution versions of geophysical datasets that comply with the FAIR principles and are suitable for programmatic access in future 2030 next-generation scalable, data-intensive computation.

Whilst there are still many unknowns about geophysical processing in 2030, we know that geophysicists will be able to work on less processed data levels and then transparently develop their own derivative products that are more tuned to the parameters of their particular use case, and they will be capable of more precise solutions that can be undertaken at regional to national scale.

In exascale environments, research teams will be able to analyse larger volumes of these high-resolution datasets, and they will be able to see the quality of their algorithms or workflows quickly and within realistic time frames. The NCI geophysics and AI/ML modules have been set up as a first try at integrating thousands of Python/Julia/R libraries and packages related to the geosciences, data science, HPC scaling and visualisation. These modules aim to be community driven, regularly updated and will hopefully ease the burden of researchers having to spend an inordinate amount of time on software engineering. Additionally, the NCI HPC software library of standalone geophysical codes that support parallel computation are being expanded and made available to the community.

As noted by Rees et al. (in press), HPC computation is not widely used by the minerals exploration industry in Australia. In contrast, the petroleum industry has embraced HPC for processing of geophysics data over the past two decades. If we can enable explorationists to analyse their datasets at higher resolution, using flexible software environments that can be more precisely targeted at local conditions, will discovery rates increase?

ACKNOWLEDGMENTS

This work has been supported by co investments from AuScope, NCI and the ARDC Cross-NCRIS National Data Assets Project XN002: National high-resolution geophysics reference collections for 2030 computation https://doi.org/10.47486/XN002.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. and Ghemawat, S., 2016, Tensorflow: Large-scale machine learning on heterogeneous distributed systems. https://doi.org/10.48550/arXiv.1603.04467

Australian Research Data Commons (ARDC), 2021, ARDC Cross-NCRIS National Data Assets Project XN002: National high-resolution geophysics reference collections for 2030 computation. https://doi.org/10.47486/XN002.

Barth, B., Foster, C., Pronk, M., et al., 2022, JuliaGeo: GitHub repository. https://github.com/JuliaGeo

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. and Wassermann, J., 2010, ObsPy: A Python toolbox for seismology: Seismological Research Letters, 81(3), pp.530-533. https://doi.org/10.1785/gssrl.81.3.530

Byrne, S., Wilcox, L.C. and Churavy, V., 2021, MPI.jl: Julia bindings for the Message Passing Interface: JuliaCon Proceedings, 1(1), 68. https://doi.org/10.21105/jcon.00068

Chave, A.D. and Thomson, D.J., 2003, A bounded influence regression estimator based on the statistics of the hat matrix: Journal of the Royal Statistical Society: Series C (Applied Statistics), 52(3), 307-322. https://doi.org/10.1111/1467-9876.00406

Cockett, R., Kang, S., Heagy, L.J., Pidlisecky, A. and Oldenburg, D.W., 2015, SimPEG: An open source framework for simulation and gradient based parameter estimation in geophysical applications: Computers & Geosciences, 85, 142-154. https://doi.org/10.1016/j.cageo.2015.09.015

Collette, A., Kluyver, T., Caswell, T., et al., 2022, h5py/h5py: 3.7.0 (3.7.0): Zenodo. https://doi.org/10.5281/zenodo.6575970

Collins, B., Van de Ven, B., Eswaramoorthy, P., Zimmermann, M., Avrella, D., 2020, Bokeh: Essential Open Source Tools for Science: Zenodo. https://doi.org/10.5281/zenodo.4317718

Dalcin, L. and Fang, Y.L.L., 2021, mpi4py: Status update after 12 years of development: Computing in Science & Engineering, *23*(4), 47-54. https://doi.org/10.1109/MCSE.2021.3083216

Danisch, S. and Krumbiegel, J., 2021, Mackie.jl: Flexible high-performance data visualization for Julia: Journal of Open Source Software, 6(65), 3349. https://doi.org/10.21105/joss.03349

Folch, A., de la Puente, J., Sandri, L., Halldórsson, B., Fichtner, A., Gracia, J., Lanucara, P., Bader, M., Gabriel, A.-A., Macías, J., Lovholt, F., Fournier, A., Monteiller, V., and Laforet, S., 2020, e-infrastructures and natural hazards.

The Center of Excellence for Exascale in Solid Earth (ChEESE): EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-13497. https://doi.org/10.5194/egusphere-egu2020-13497

Gabriel, E., Fagg, G.E., Bosilca, G., Angskun, T., Dongarra, J.J., Squyres, J.M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A. and Castain, R.H., 2004, Open MPI: Goals, concept, and design of a next generation MPI implementation: In European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting (pp. 97-104), Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30218-6_19

Granger, B.E. and Pérez, F., 2021, Jupyter: Thinking and storytelling with code and data: in Computing in Science & Engineering, 23(2), 7-14. https://doi.org/10.1109/MCSE.2021.3059263

Gross, L., Bourgouin, L., Hale, A.J. and Mühlhaus, H.B., 2007, Interface modeling in incompressible media using level sets in Escript: Physics of the Earth and Planetary Interiors, 163(1-4), 23-34. https://doi.org/10.1016/j.pepi.2007.04.004

Heimann, S., Kriegerowski, M., Isken, M., Cesca, S., Daout, S., Grigoli, F., Juretzek, C., Megies, T., Nooshiri, N., Steinberg, A. and Sudhaus, H., 2017, Pyrocko - An open-source seismology toolbox and library: GFZ Data Services, https://doi.org/10.5880/GFZ.2.1.2017.001

Hoyer, S. and Hamman, J., 2017, Xarray: ND labeled arrays and datasets in Python: Journal of Open Research Software, *5*(1). http://doi.org/10.5334/jors.148

Ip, A., Turner, A., Poudjom-Djomani, Y., Brodie, R., Wynne, P., Druken, K., Symington, N., and Kemp, C., 2019, Discovering and using geophysical data in the 21st century: ASEG Extended Abstracts, 2019:1, 1-6. https://doi.org/10.1080/22020586.2019.12073191

Jordahl, K., Bossche, J., Fleischmann, M., et al., 2021. geopandas/geopandas: v0.10.2 (v0.10.2): Zenodo. https://doi.org/10.5281/zenodo.5573592

Kelbert, A., 2020, EMTF XML: New data interchange format and conversion tools for electromagnetic transfer functions: GEOPHYSICS, 85, F1-F17. https://doi.org/10.1190/geo2018-0679.1

Kelbert, A., Meqbel, N., Egbert, G.D. and Tandon, K., 2014, ModEM: A modular system for inversion of electromagnetic geophysical data: Computers & Geosciences, 66, 40-53. https://doi.org/10.1016/j.cageo.2014.01.010

Key, K., 2016, MARE2DEM: a 2-D inversion code for controlled-source electromagnetic and magnetotelluric data: Geophysical Journal International, 207(1), 571-588. https://doi.org/10.1093/gji/ggw290

Kirkby, A.L., Zhang, F., Peacock, J., Hassan, R. and Duan, J., 2019, The MTPy software package for magnetotelluric data analysis and visualisation: Journal of Open Source Software, 4(37), 1358. https://doi.org/10.21105/joss.01358

Krieger, L. and Peacock, J.R., 2014, MTpy: A Python toolbox for magnetotellurics: Computers & geosciences, 72, 167-175. https://doi.org/10.1016/j.cageo.2014.07.013

Miles, A., Kirkham, J., Bussonnier, M., et al., 2022, zarr-developers/zarr-python: v2.11.3 (v2.11.3): Zenodo. https://doi.org/10.5281/zenodo.6419641

Moritz, P., Nishihara, R., Wang, S., et al., 2018, Ray: A distributed framework for emerging {AI} applications: In 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18) (pp. 561-577). Retrieved November 14, 2022 from https://www.usenix.org/system/files/osdi18-moritz.pdf

Pagani, M., Monelli, D., Weatherill, G., Danciu, L., Crowley, H., Silva, V., Henshaw, P., Butler, L., Nastasi, M., Panzeri, L. and Simionato, M., 2014, OpenQuake engine: An open hazard (and risk) software for the global earthquake model: Seismological Research Letters, 85(3), 692-702. https://doi.org/10.1785/0220130087

Papadopoulos, S., Datta, K., Madden, S. and Mattson, T., 2016, The tiledb array data storage manager: Proceedings of the VLDB Endowment, 10(4), 349-360. https://doi.org/10.14778/3025111.3025117

Paszke, A., Gross, S., Massa, F., et al., 2019, PyTorch: An Imperative Style, High-Performance Deep Learning Library: In Advances in Neural Information Processing Systems 32 (pp. 8024–8035), Curran Associates, Inc. Retrieved November 14, 2022 from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Peacock, J., Kappler, K., Heagy, L., Ronan, T., Kelbert, A. and Frassetto, A., 2022a, MTH5: An archive and exchangeable data format for magnetotelluric time series data: Computers & Geosciences, 162, 105102, https://doi.org/10.1016/j.cageo.2022.105102

Peacock, J., Kappler, K., Keyson, L. Heagy, L., and Rees, N., 2022b, kujaku11/mth5: v0.3.0 (v0.3.0): Zenodo. https://doi.org/10.5281/zenodo.7111888

Peacock, J.R., Frassetto, A., Kelbert, A., Egbert, G., Smirnov, M., Schultz, A.C., Kappler, K.N., Ronan, T., and Trabant, C., 2021, Metadata Standards for Magnetotelluric Time Series Data: U.S. Geological Survey data release. https://doi.org/10.5066/P9AXGKEV

Perez, F. and Granger, B.E., 2015, Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science. Retrieved November 11, 2022 from https://blog.jupyter.org/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011, Scikit-learn: Machine learning in Python: The Journal of machine Learning research, 12, 2825-2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Plotly Technologies Inc., 2015, Collaborative data science: Montréal: Plotly Technologies Inc Montréal, Available online: https://plot.ly (accessed on 14 November 2022).

Puzyrev, V., 2019, Deep learning electromagnetic inversion with convolutional neural networks: Geophysical Journal International, 218(2), 817–832, https://doi.org/10.1093/gji/ggz204

Rathgeber, F., Ham, D.A., Mitchell, L., Lange, M., Luporini, F., McRae, A.T., Bercea, G.T., Markall, G.R. and Kelly, P.H., 2016, Firedrake: automating the finite element method by composing abstractions: ACM Transactions on Mathematical Software (TOMS), 43(3), 1-27. https://doi.org/10.1145/2998441

Ray, A., Taylor, R., Moghaddam, N., and Symington, N., 2022, HiQGA: GitHub repository. https://github.com/GeoscienceAustralia/HiQGA.jl

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 High Performance Computing Platform: ASEG Extended Abstracts, 2019(1), 1-6. https://doi.org/10.1080/22020586.2019.12073015

Rees, N., Wang, S., Evans, B., Wyborn, L., Rawling, T., Goleby, B., Druken, K., and Yang, R., in press, Using the NCI Gadi Supercomputer to revolutionise processing of MT time series data: results from the GeoDeVL experiment: ASEG 2021 extended abstracts.

Rocklin, M., 2015, Dask: Parallel computation with blocked algorithms and task scheduling: In Proceedings of the 14th python in science conference (Vol. 130, p. 136), Austin, TX: SciPy. Retrieved November 14, 2022 from https://conference.scipy.org/proceedings/scipy2015/pdfs/matthew_rocklin.pdf

Rouault, E., Warmerdam, F., Schwehr, K., et al., 2022, GDAL (v3.4.2): Zenodo. https://doi.org/10.5281/zenodo.6352176

Schaa, R., Gross, L. and Du Plessis, J., 2016, PDE-based geophysical modelling using finite elements: examples from 3D resistivity and 2D magnetotellurics: Journal of Geophysics and Engineering, 13(2), S59-S73. https://doi.org/10.1088/1742-2132/13/2/S59

Sergeev, A. and Del Balso, M., 2018, Horovod: fast and easy distributed deep learning in TensorFlow. Retrieved on November 14 2022 from https://arxiv.org/abs/1802.05799v3

Usui, Y., 2015, 3-D inversion of magnetotelluric data using unstructured tetrahedral elements: applicability to data affected by topography: Geophysical Journal International, 202(2), 828-849. https://doi.org/10.1093/gji/ggv186

Waskom, M.L., 2021, Seaborn: statistical data visualization: Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Wickham, H., François, R., Henry, L., Müller, K., 2022, dplyr: A Grammar of Data Manipulation. Retrieved on November 14 2022 from https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Wickham, H. and Girlich, M., 2022, tidyr: Tidy Messy Data. Retrieved on November 14 2022 from https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York. ISBN 978-3-319-24277-4. Retrieved on November 14 2022 from https://ggplot2.tidyverse.org.

Whitaker, J., Khrulev, C., Huard, D., et al., 2020, Unidata/netcdf4-python: version 1.5.5 release (v1.5.5rel2): Zenodo. https://doi.org/10.5281/zenodo.4308773

Wight, D.E., 1988, SEG standard for MT and EMAP data: In SEG Technical Program Expanded Abstracts 1988 (pp. 249-251), Society of Exploration Geophysicists. https://doi.org/10.1190/1.1892244

Wilkinson, M.D, Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., and Bouwman, J., 2016, The FAIR Guiding Principles for scientific data management and stewardship: Scientific data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Witte, P.A., Louboutin, M., Kukreja, N., Luporini, F., Lange, M., Gorman, G.J. and Herrmann, F.J., 2019, A large-scale framework for symbolic implementations of seismic inversion algorithms in Julia: Geophysics, 84(3), F57-F71. https://doi.org/10.1190/geo2018-0174.1

Wrona, T., Pan, I., Bell, R.E., Gawthorpe, R., Fossen, H., Brune, S., 2021. 3D seismic interpretation with deep learning: A brief introduction: The Leading Edge, 40(7). https://doi.org/10.1190/tle40070524.1

Wyborn, L., Rees, N., Evans, B., Wang, S., Drucken, K., Yang, R., Guo, J., Quarat, T., Wang, J., Goleby, B., Salmon, M., Pickle, R., Klump, J., Woodman, S., Friedrich, C., Fazio, V., Fraser, R., Rawling, T., Martin, J., and Benn, J., 2020, GeoDeVL Project Final Report: Zenodo. https://doi.org/10.5281/zenodo.4278997

Wyborn, L., Rees, N., Klump, J., Evans, B., Rawling, T., and Druken, K., 2022. The Known Knowns, the Known Unknowns and the Unknown Unknowns of Geophysics Data Processing in 2030: EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-11012. https://doi.org/10.5194/egusphere-egu22-11012

Yu, S. and Ma, J., 2021, Deep learning for geophysics: Current and future trends: Reviews of Geophysics, 59(3). https://doi.org/10.1029/2021RG000742

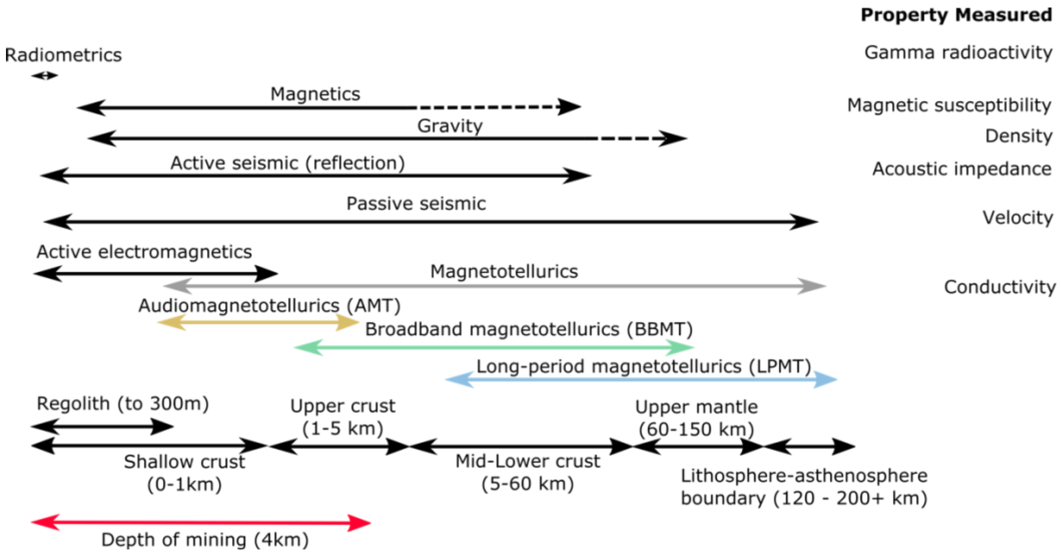

Figure 1. Types of geophysical data collected in Australia, the physical property measured and the depth of the crust that is sampled: also shown is the depth of current mining. Figure modified from original of Richard Chopping (GSWA).