Capturing (via automation) the Sequential Processing Levels along multiple Full-paths of Magnetotellurics Data Use

Nigel Rees1, Ben Evans2, Dennis Conway3, Hoël Seillé4, Bruce Goleby5, Lesley Wyborn6

1National Computational Infrastructure, Australian National University, Canberra, Australia, nigel.rees@anu.edu.au
2National Computational Infrastructure, Australian National University, Canberra, Australia, ben.evans@anu.edu.au
3École nationale supérieure d'ingénieurs de Caen, Caen, France, dennis.conway@ensicaen.fr
4CSIRO, Deep Earth Imaging Future Science Platform, Kingston, Australia, hoel.seille@csiro.au
5OPM Consulting, Canberra, ACT, Australia, bruce.goleby@opmconsulting.com.au
6National Computational Infrastructure, Australian National University, Canberra, Australia, lesley.wyborn@anu.edu.au

SUMMARY

With easier access to larger computational capacity, as well as a greater demand for transparency in science, there is a growing need amongst the Australian and international Magnetotellurics (MT) community for original field data and less processed time series data to be formally published and made more readily available online in a FAIR manner to both facilitate data reuse and repurposing and engender trust in the higher-level data products. Exposing the Full-path of MT data use [1] and being able to track back to the original data acquisition parameters, including the instrument used and any field metadata is also critical for enabling transparency and due diligence assessments of the more highly processed transfer functions and model outputs.

ACCESSING THE FULL-PATH OF MT DATA

The Full-path of any data use extends from data capture, data access and management, data analysis and modelling, through data and model intercomparisons together with data provenance systems. For MT, the different processing levels along the Full-path of data use, from the original field data acquisition to the manicured (edited, calibrated, resampled) time series to processed transfer functions to model inputs to model outputs are listed in Table 1 ([2, 3]). Each level needs to be distinguished and captured in a reproducible manner so that researchers can confidently utilise MT data at any stage they desire, as the provenance of each level can be accessed.

In Australia there already exists a large volume of publicly funded MT time series datasets and even greater volumes will be collected into the future. However, very little of this lower level time series is discoverable or accessible online, and it is usually only distributed on hard drives: the focus of MT research and publications is biased towards the more highly evolved data products.

RATIONALE

Having the less processed MT time series data accessible will facilitate future re-analysis and analytical reproducibility, validation of statistical models and comparisons of findings. Additionally, there are many potential data processing advantages that could be investigated including enhancing data quality through data selection; better understanding of noise sources; computation of the full error covariance of the impedance estimates; assessment of results using different approaches and codes; and merging of multiple surveys, particularly when running national scale analysis. To leverage these opportunities, time series data needs to be more accessible and comply with the FAIR principles, preferably co-located next to HPC compute so that as data is generated, it can be easily captured, ingested and processed or re-processed rapidly.

Inevitably there will be multiple versions of each level of the MT data and therefore automations must be developed so that each level of MT processing along multiple paths can be carefully constructed from the previous in a reproducible manner. Fundamental to these automations are international agreements on what data formats and vocabularies are to be used, and what metadata is required at each level.

A PROPOSAL

The Australian Research Data Commons (ARDC) funded Geoscience Data Enhanced Virtual Laboratories (GeoDeVL) project is working towards stabilising, automating and increasing the consistency and quality of the different data levels and products in the MT data life cycle. It is important to recognise that it is quite likely that each level could have different teams producing it. By separating and applying a persistent identifier and creating a landing page for each sequential layer it is possible to attribute all those individuals and organisations that were involved in creating each layer, as well as those that provided funding. In particular, for the level 0 layers, the specific individual instrument that was used to acquire the data can be recorded, as well as all those who contributed anything to collecting the data in the field, including those that dug the holes!

Table 1: The sequential levels of MT data processing (from [2,3])

Processing Levels	Name	Description	Collection / Processed By	Typical Volumes
Packed Raw Data	Raw Time Series	Original field data streamed from site data loggers	Single researcher or research team	GBs to TBs
Level 0	Edited Time Series	Time ordered instrument recorded data (e.g., raw voltages, counts) at full resolution	Single researcher or research team	GBs to TBs
Level 1A	Calibrated Time Series	Level 0 data that have been calibrated in a reversible manner and packaged with associated calibration equations	Single researcher or research team	GBs to TBs
Level 1B	Resampled Time Series	Level 0 or 1A data that have been irreversibly transformed (e.g., resampled, noisy data removed, filters applied)	Can be processed by anyone with access to L1A	GBs to TBs
Level 2	Derived Frequency Domain Processed Data (e.g., EDI)	Geophysical parameters (e.g., impedance tensors) derived from frequency domain time series processing of Level 1A or 1B data	Can be processed by anyone with access to L1A or L1B	MBs
Level 3A	Derived modelling inputs	Level 2 parameters converted into input files for modelling and inversion algorithms	Can be processed by anyone with access to L2	MBs
Level 3B	Derived modelling outputs	Level 2 parameters mapped onto space-time grids	Can be processed by anyone with access to L2 or L3A	MBs

REFERENCES

1. Asch, M., et al., Big data and extreme-scale computing:Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry. The International Journal of High Performance Computing Applications, 2018. 32(4): p. 435–479. https://doi.org/10.1177/1094342018778123, Accessed 24 September 2019.
2. Rees, N., et al., 2019, The Geosciences DeVL Experiment: new information generated from old magnetotelluric data of The University of Adelaide on the NCI High Performance Computing Platform. AEGC 2019 Data to Discovery, August 2019, Perth.
3. NASA EARTHDATA. 2019. Data Processing Levels. Available from: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels, Accessed 25 September 2019.

ACKNOWLEDGMENTS

The Geosciences DeVL project was funded by ARDC and AuScope with in kind support from NCI, CSIRO and The University of Adelaide. The authors also acknowledge support from the Australian Government Department of Education, through the National Collaboration Research Infrastructure Strategy (NCRIS), and the National Computational Infrastructure (NCI) including its partners (ANU, CSIRO, BoM and GA) that contributed to building the data and compute infrastructures which enabled HPC techniques to be implemented on MT data at NCI. The DeVL project builds on trials with MT on HPC from 2012 to 2016 between GA and NCI.