GeoDeVL Project Final Report

project title	Geoscience Data Enhanced Virtual Laboratory
project number	
Contracting organisation	AuScope
date	15 October 2020
prepared by	Lesley Wyborn
co-authors/ contributors/ collaborators	Nigel Rees, Ben Evans, Sheng Wang, Kelsey Drucken, Rui Yang, Jian Guo, Qurat Tariq, Jingbo Wang (NCI); Bruce Goleby (OPM Consulting); Michelle Salmon, Robert Pickle (RSES, ANU); Jens Klump (AuScope and CSIRO); Stuart Woodman, Carsten Friedrich, Vincent Fazio (CSIRO); Ryan Fraser (CSIRO, (now AARNet)); Tim Rawling (AuScope); Julia Martin, Joel Benn (ARDC).
approved by	

Contents

1. Acknowledgments

For the **MT Work Package,** we note that there were no pre-existing public domain examples of what we were trying to do with the online publication of MT time series data and in the application of HPC to Level 0 and Level 1 MT data processing. The GeoDeVL project team would therefore particularly 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, Tim Ahern, Jerry Carter, Bob Woodward and Rob Casey (IRIS, USA); Jared Peacock, Anna Kelbert (USGS); and Kirsten Elger (GFZ, Potsdam).

The **Passive Seismic Work Package** would like to acknowledge the help of Megan Miller (Research School of Earth Sciences, ANU); Chad Trabant, Tim Ahern, Jerry Carter (IRIS); Melanie Barlow (ARDC); Irina Bastrakova (ANZLIC Metadata Working Group and Geoscience Australia); and Kirsten Elger (GFZ Potsdam).

The **IGSN Work Package** would like to acknowledge the help of Kerstin Lehnert and Sarah Ramdeen (Columbia University, USA).

The **AVRE Work Package** would like to acknowledge Peter Warren, Carsten Laukamp (CSIRO); and members of the Geophysical Acquisition and Processing Team and Alex Ip (Geoscience Australia) in providing help with the versions of the ASTER and the National Geophysical datasets that NCI then published and made accessible to the AuScope and VGL Portals via the GSKY server.

2. Executive summary

In response to the 2016 National Research Infrastructure Roadmap[1] 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. AuScope Virtual Research Environments (AVRE)[2] is designed to support the DLT as a highly flexible online 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 high performance supercomputers. The key requirements for data in AVRE are that:

1. Relevant Solid Earth data from research, government and industry sources are Findable, Accessible, Interoperable and Reusable (FAIR) across three coordinated research data networks - **Geophysics**, **Geochemistry** and **Geology**; and
2. In line with the AuScope 10 year Strategy for 2020-2030[3], AVRE supports predictive geoscience by applying global data principles and data management best practice.

Since 2008, as part of the National Collaborative Research Infrastructure Strategy (NCRIS), there has been a long history of investments by AuScope, NCI, ARDC (formerly ANDS, NeCTAR, RDS) and government agencies in online data infrastructures. A review in 2017 showed that investments to date had been focussed more on access to geological datasets of the government agencies, and that there had been little effort on Geophysics and Geochemistry research infrastructures. Existing infrastructures were dated, e.g. the Virtual Laboratory Infrastructure was based around workflow engines and predated Python/Jupyter notebooks. Hence, the GeoDeVL program was seen as an opportunity to both accelerate the development of the Geophysics and Geochemistry networks and to foster a change towards a more flexible, virtual research environment that enabled researchers to more easily compose their own workflows specifically tailored to their specific research questions. Work Packages 1 and 2 focussed on making AuScope funded Magnetotelluric (MT) and Passive Seismic (PS) geophysical data available online whilst Work Package 3 focussed on the development of an ability to use IGSNs

[1] https://docs.education.gov.au/system/files/doc/other/ed16-0269_national_research_infrastructure_roadmap_report_internals_acc.pdf
[2] https://www.auscope.org.au/avre
[3] https://static1.squarespace.com/static/5b440dc18ab722131f76b631/t/5f3e111d0cfa573f4f3eac39/1597903187392/10-Year+Strategy+2020+%E2%80%93+2030+%E2%80%94+Online.pdf

to uniquely identify samples analysed in the recently launched AuScope Geochemistry Network[4]. Work Package 4 focussed on building the foundations of AVRE and was a balance between adding new systems and at the same time, modernizing and reducing the growing technical debt of the earlier infrastructure investments.

The objectives and key highlights from each work package are as follows:

1) In the **Magnetotellurics Work Package** a transparent workflow was developed for processing the rawer, full resolution MT data collected with Earth Data Logger instrument: by using HPC, data processing times were reduced from days and weeks to minutes. Researchers no longer have to use higher level, derivative data products/ models developed to parameters that data providers choose: instead they can now develop their own products transparently. An additional bonus is that as new processing methodologies and/or higher capacity computers become available, the rawer forms of earlier surveys are available online and can easily be accessed, reprocessed and more easily adapted to new surveys.
2) With the GeoDeVL and other funding sources, the **Passive Seismic Work Package** made PS data from surveys run over the last 30 years accessible online using the International Federation of Digital Seismograph Networks (FDSN) standards and protocols. The GeoDeVL project focussed on crosswalking the FDSN metadata into more generic standards so that PS seismic data can be discovered along with other geophysical data types in national portals and catalogues.
3) The **IGSN Work Package** was focussed around increasing uptake of IGSN in departmental laboratories. Unfortunately, due to a combination of a major overhaul of the recommended IGSN architecture internationally and the lack of viable supporting infrastructures in most Earth science departments it was not possible to complete this.
4) The **Australian Virtual Research Environments Work Package** incorporated the new published datasets from data services such as GSKY at NCI: some processing systems were updated and some legacy systems were modernised.

3. Background

The Australian Research Data Commons (ARDC) co-funded the Geosciences DeVL Project over three phases, and was seen as an ideal opportunity to kick-start AuScope's 2017 intention for an integrated 5-year program to develop the AuScope Virtual Research Environment (AVRE), a coordinated eResearch Infrastructure that will enable researchers to combine analytical methods such as inversions and simulations, as well as real-world observations to develop probabilistic descriptions of specific Earth processes.

A key driver for the creation of AVRE was the 2016 National Roadmap for Research Infrastructure (NRIR), which maintained that in order to secure global leadership for Australian Earth sciences over the next decade "*there is now a need to enhance integration of existing data and mathematical modelling across large geographical areas to establish the next generation* of

[4] https://www.auscope.org.au/posts/agn

'inward looking telescopes' to better understand the evolution of Earth's crust and the resources contained within it".

The new AVRE ambition firstly requires data to be Findable, Accessible, Interoperable and Reusable (FAIR) across three coordinated Earth science data networks: **Geophysics**, **Geochemistry** and **Geology** with a goal of enabling 4D analysis of the Australian continent. At this scale, data will need to be both human readable and machine actionable, which in turn requires adherence to agreed community standards both within and beyond Solid Earth Sciences. Secondly, it requires online access to tools and software for easy, on-the-fly data visualisation of the full (3D) datasets, and analysis to enhance and accelerate knowledge generation. Finally, users require access to a variety of computational infrastructures that range from high-end HPC, to cloud (both public or private) to local hand-held devices.

Prior to the start of the GeoDeVL project, most of the AuScope investments on data had been on the **Geological** Data Networks which were built more in partnership with Geoscience Australia, and the State and Territory Geological Surveys. The 2018-2020 Geosciences DeVL project intended to accelerate the development of the Australian **Geophysical** and **Geochemical** Research Data Networks, which linked with equivalent international research infrastructures and networks, and where possible linked to equivalent networks being developed by the Australian Government Geological Surveys.

Each GeoDeVL project was run as 4 separate, but related work packages:

1. A Magnetotelluric (MT) Work Package led by NCI, The University of Adelaide and the Geological Survey of South Australia;
2. A Passive Seismic Work Package led by RSES, ANU;
3. An IGSN Work Package led by CSIRO and the Australian Research Data Commons (ARDC);
4. The integrative AVRE work package led by NCI and CSIRO.

The first three of these work packages targeted research data types that at best in 2017 were stored in offline repositories and could not be discovered online. The fourth package aimed to create the foundations for AVRE by building on, and better coordinating existing Australian geoscience eResearch infrastructures that have been developed over the last decade in data, tools and Virtual Laboratories (VLs) through funding from AuScope, National eResearch Collaboration Tools and Resources (NeCTAR), Australian National Data Service (ANDS), Research Data Storage Infrastructure (RDSI), Research Data Services (RDS), NCI, CSIRO, GA, the State and Territory Geological Surveys including the existing:

1. NeCTAR/ANDS-funded Virtual Geophysics Laboratory (VGL);
2. ANDS funded Trusted research outputs for software - the Scientific Software Solutions Centre (SSSC);
3. The RDS A1.9a Mineral Data Services and A1.9b Geophysics data services projects;
4. AuScope Investments in eResearch Infrastructure since 2006; and
5. National Collaborative Research Infrastructure Strategy (NCRIS) and partner investment in the National Computational Infrastructure (NCI) Facility.

Most of these projects have had a long history and hence there was an urgent need to update them; some of them substantially. For example, the GeoDeVL project has its origins in the first Virtual Geophysics Laboratory (VGL)[5] that was funded as part of the NeCTAR Virtual Laboratory 2012 Early Adopter Program. VGL in turn was based on components that were prototyped in the 2009-2011 ANDS funded Australian Spatial Research Data Commons Project[6]. Clearly, much has changed since then - in particular, Python/Jupyter notebooks did not exist at the start and early versions of VGL relied heavily on workflow design, which were reasonably flexible, but still had limitations that led to increased costs and sustainability issues. Where notebooks have now been introduced into AVRE components, their focus has been more on enabling training. The goal is to help finalise a shift to researchers being able to find data and tools as online services and then compose their own workflows to suit their specific research needs.

The outcomes of this final GeoDeVL extension were focussed on:

1. Increasing the transparency and diversity of MT data products by firstly exposing the rawer forms of MT data and then secondly using HPC to automate and optimise the processing and publishing of the Archived, Level 0 and Level 1 time series MT data;

2. Enabling better interaction of the AuScope funded eResearch systems with infrastructure designed to be used by the whole research system. This is intended to be a pathway for Earth Science data to be used in interdisciplinary science (e.g., connectivity of the MT and PS data communities into Research Data Australia (RDA[7]); enabling tracking of datasets used in publications through tools such as SCHOLIX[8], Research Graph[9]; and ensuring AuScope IGSN minted samples are compatible with other samples in other domains, as well as across the research, government and CSIRO communities);

3. Investigating how to better support the Australian research community in the implementation of IGSN at the level of the individual field-based researcher and/or small departments with in situ instrument laboratories; and

4. Creating a one-stop-shop - the AuScope Virtual Research Environments (AVRE) - that enable access to a rich ecosystem of Findable, Accessible, Interoperable and Reusable (FAIR, Wilkinson et al., 2016) data and tools that are of value to the research community and come from a diverse range of Australian research organisations, Geological Surveys, industry and the international community. AVRE focuses on facilitating real time data assimilation and analysis to address both simple and complex problems at any scale.

4. Objectives

The objectives of the third and final phase of the GeoDeVL project were:

1) ***MT Work Package: make two MT time series AusLAMP surveys accessible from the NCI***

[5] https://nectar.org.au/labs/virtual-geophysics-laboratory/
[6] https://projects.ands.org.au/id/EIF003
[7] https://researchdata.edu.au/
[8] http://www.scholix.org/
[9] https://researchgraph.org/

Catalogue and Research Data Australia compatible with international standards through:

a) Stabilising and increasing the consistency and QA/QC of the MT data publishing pipelines starting with the rawest forms of the data, with an aim of increasing transparency and reproducibility of MT data: trial MT HPC performance improvements;
b) Update the University of Adelaide Musgraves and University of Tasmania AusLAMP datasets to be compliant with the agreed Community metadata standards; and
c) Continue to ensure all new MT datasets added to the NCI Data platform as part of the GeoDeVL extension project are discoverable in the NCI catalogue, the data made usable through services and integrated into the AuScope/VGL Portals, and metadata harvested in the RDA catalogue.

2) ***Passive Seismic Work Package: report on potential for AusPass data to be accessible in Australian Research Data catalogues and portal***
   a) Determine how to make AusPass Passive Seismic data accessible in RDA.

3) ***IGSN Work Package: Investigating how to better support the Australian Research community in the Implementation of IGSN***
   a) In collaboration with the AuScope Portal, develop a searchable map of all IGSN samples registered in Australia by Australian Allocating agents; and
   b) Investigate the most sustainable way of machine to machine bulk uploading of IGSN datasets from individual geoscientific laboratories, researchers and departments.

4) ***AVRE Work Package - continue modernisation of existing AVRE components by upgrading several important components to new technologies and where possible, update to more recent versions of the datasets as they become available.***
   a) Update existing geophysics datasets to the new NCI OGC Compliant GSKY services;
   b) Developing training notebooks to convert old VGL and other workflows into notebook technology.

5. Methodology

Work Package 1: Magnetotellurics

This package aimed to make metadata/data of the rawer MT time series compliant with the FAIR principles using the AusLAMP Musgraves and Tasmania datasets. After the project started, GA informed us they were publishing the Tasmanian dataset and no further work was done on it.

For the Musgraves MT time series datasets, the data standards and processing methods were found to be immature when compared with other techniques such as passive seismic and gravity. MT geophysicists use multiple disconnected systems to acquire, store and process data, starting at the initial raw, full resolution data data collected in the field (Level 0) and the initial data editing and calibration (Level 0, Level 1 - both still at full resolution), through to the derivation of data products (Levels 2 and 3) (See Figure 1; Rees et al., 2019).

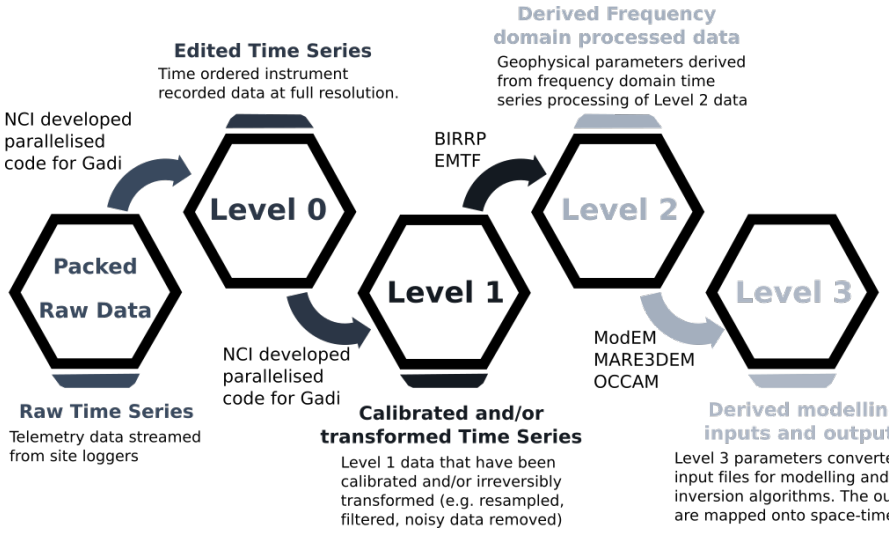


Figure 1: Emphasis of the GeoDeVL extension was the Level 0 and Level 1 datasets. The previous GeoDeVL projects focused on the Level 2 products. See Rees et al. (2019) for an explanation of Levels of processing as applied to MT.

In late 2019, a review of standards used by the International/National MT Community found multiple standards and formats were being used/being developed/being proposed (see Appendix 2 for details). In summary, at the start of the GeoDeVL extension project there was:

- Limited online access to the rawer forms of any MT data: at best time series data was only accessible via personal contact with the author/organisation that collected the data;
- No single agreed mechanism for capturing MT (meta)data at any processing level;
- No internationally agreed vocabularies for MT (meta)data which hinders the development of QA/QC protocols that ensure seamless, programmatic access to the data (all levels);
- No exemplars anywhere on making time series data accessible online; and
- No consistent processing pipeline that informs how new MT (meta)data are acquired, curated, stored and processed (Levels 0, 1);

There were essentially two alternative metadata standards for MT time series data: neither were complete at the time this work started. The project initially tried using the draft Australian metadata standard proposed to the Australian Society of Exploration Geophysicists (ASEG) by Kirby et al. (2019), but for machine actionable modes there were some issues (e.g., ambiguities including backslashes, capital letters and spaces in the variable names). We did not have these issues with the NSF-funded Incorporated Research Institutions for Seismology (IRIS) ElectroMagnetic Advisory Committee (EMAC) MT standard, which meets the requirements of the AuScope 10 year Strategy for 2020-2030 for the application of global data principles and data management best practice. The EMAC standard covers AudioMagnetoTelluric (AMT), Broadband (BB) and Long Period (LP) data but this was not released until July 2020 and has not yet been endorsed by the Australian community. There was also uncertainty over which of the Hierarchical Data Format (HDF)-based high performance data formats to use, with some Australians favouring the Adaptable Seismic Data Format (ASDF), whilst international MT researchers seem to prefer MTH5. GA is currently working with Intrepid Geophysics to survey the most appropriate standard for Australia for BB MT data, but the results from this work are not yet available, and as the GeoDeVL project was based on LP MT data they may not be applicable anyway.

As noted in the project proposal, the default option for the project was to use netCDF to store time series data if the data format issue was not resolved. The archived, Level 0 and Level 1 Musgrave time series datasets were published on the NCI THREDDS Data server and are discoverable through the NCI geonetwork catalogue[10,11]. Metadata on these datasets has also been harvested into the AuScope[12] and VGL[13] portals and RDA.

The GeoDeVL project automated and optimised the processing and publishing of the archived, Level 0 and Level 1 time series data from Earth Data Loggers instruments. Considerable speed performances were achieved: processing that previously took days and weeks, was reduced to minutes (Table A3.1 in Appendix 3 lists time comparisons). A presentation[14] is available which elaborates on the methodology used and results. However, this method is still developmental and will continue to be so until the Australian community stabilises on which standards and formats they are going to use.

The **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**MTtsdp**) codes[15] were developed to enable others to prepare their MT time series meta(data) as Archived, Level 0 and Level 1 products. These codes are unique to the type of time series recording instrument used and were developed for the Earth Data Loggers funded by AuScope, and available from the ANSIR Research Facilities for Earth Sounding[16] instrument pool. The USGS has produced similar codes for the Zonge International developed ZEN data loggers[17]. There are other MT time series logger instruments deployed in Australia (e.g., LEMI, Phoenix) and an equivalent set of codes would need to be developed per instrument type.

This tying of processing to the specific instrument type used to collect the data attests to the immaturity of the MT standards, both internationally and nationally, and contrasts strongly with the global Passive Seismic community which has stabilised on the FDSN standards and protocols. There is discussion in the US about making the new EMAC MT standards into an 'FDSN equivalent', but that will be sometime into the future. Until an MT standard is adopted internationally, processing of MT data at the lower levels will be complex, instrument specific and hard to standardise.

Work Package 2: Passive Seismic

ANU AusPass metadata was previously harvested into the AuScope portal via the NSF-funded IRIS Data Management Center[18] in Seattle using the FDSN web service specification[19] in an embarrassingly convoluted process, whereby Australian metadata was harvested from the US

[10] http://dapds00.nci.org.au/thredds/catalogs/my80/AusLAMP/AusLAMP.html
[11] http://dx.doi.org/10.25914/5eaa30cc934d0
[12] http://portal.auscope.org.au/
[13] https://vgl.auscope.org/data
[14] https://docs.google.com/presentation/d/1MJfp20AlTBID9WbuU8UJXNHDdvXvKHwv53aWRA1OuS8/edit#slide=id.g8254bb5273_0_31
[15] https://github.com/nci/MTtsdp
[16] http://ansir.org.au/
[17] https://github.com/kujaku11/MT_Zen_Master
[18] https://ds.iris.edu/ds/nodes/dmc/
[19] https://www.fdsn.org/webservices/fdsnws-availability-1.0.pdf

IRIS data center in Seattle in FDSN compliant formats, converted into an old ISO 19115 metadata profile and then made accessible from the AuScope portal (Figure 2) and then harvested by RDA.

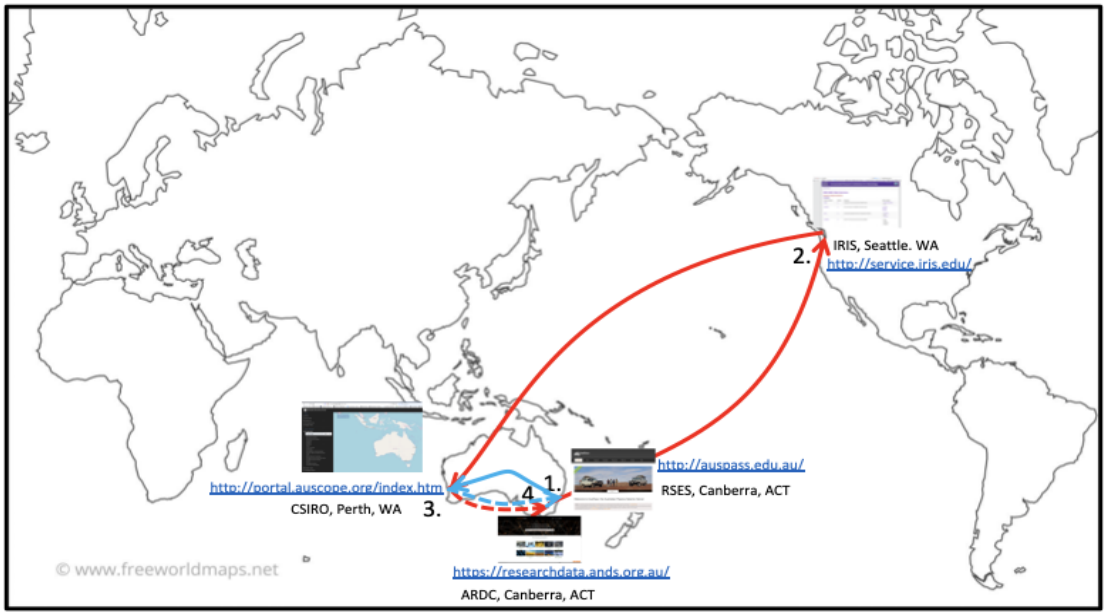


Figure 2: Previous harvesting path in red of the AusPass metadata from RSES (ANU) to IRIS to AuScope portal to RDA. New path in blue is direct from ANU to AuScope. The link to RDA from the AuScope portal needs updating.

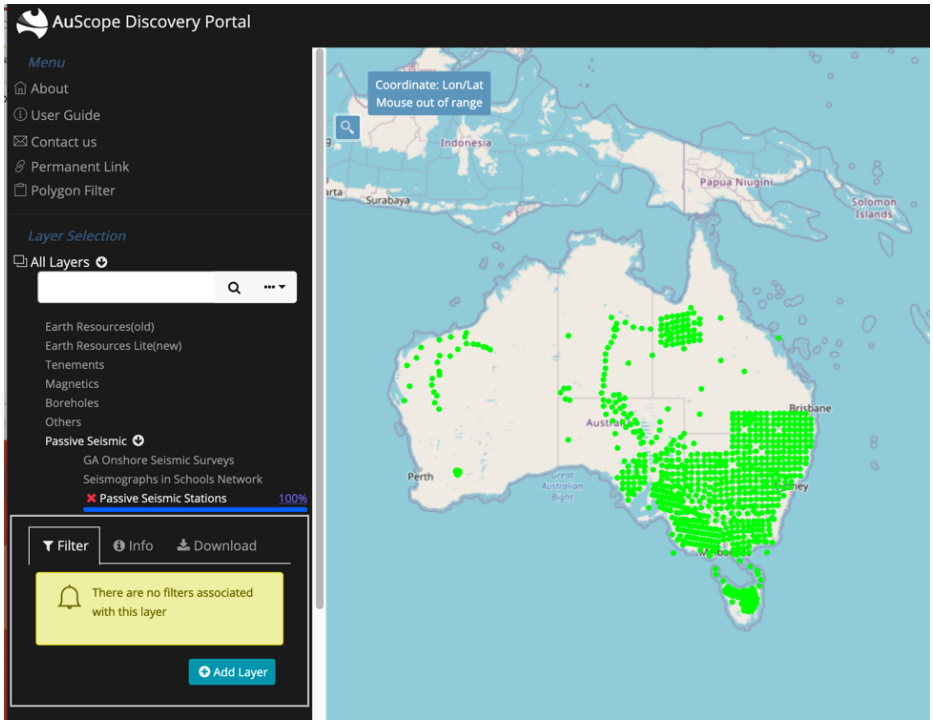


Figure 3: The AusPass Metadata is now discoverable in the AuScope portal.

The AuScope portal[20] now accesses AusPass metadata directly from ANU (Figure 3), but because the AuScope portal currently uses an old version of GeoNetwork and an old profile of ISO 19115, the AusPass metadata cannot be harvested by RDA. AuScope is currently upgrading both

[20] http://portal.auscope.org.au/

GeoNetwork and the ISO 19115 profile and this will reestablish the link with RDA. Alternatively, if a metadata crosswalk is written from FDSN, RDA could harvest directly from AusPass. This issue is currently being investigated with the help of ARDC and the ANZLIC Metadata Working Group.

Work Package 3: IGSN

For **Output 1** of the IGSN Work Package, the GeoDeVL project tried to meet demand from the community for a portal that showed the location of Australian IGSN registered sites. The dedicated portal for IGSN data, previously developed by AuScope, was deprecated by this project as it was technologically too obsolete. An attempt was made using the main AuScope Portal to access to Australian IGSN sites using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [21] which is a low-barrier mechanism for repository interoperability whereby repositories expose structured metadata via OAI-PMH, and Service Providers then make OAI-PMH service requests to harvest that metadata. The four IGSN allocating agents were:

1. CSIRO: https://igsn.csiro.au/csiro/service/oai
2. ARDC: https://handle.ands.org.au/igsn/api/service/30/oai
3. Geoscience Australia (GA): http://pid.geoscience.gov.au/sample/oai
4. GFZ Potsdam: http://doidb.wdc-terra.org/igsnaaoaip/oai

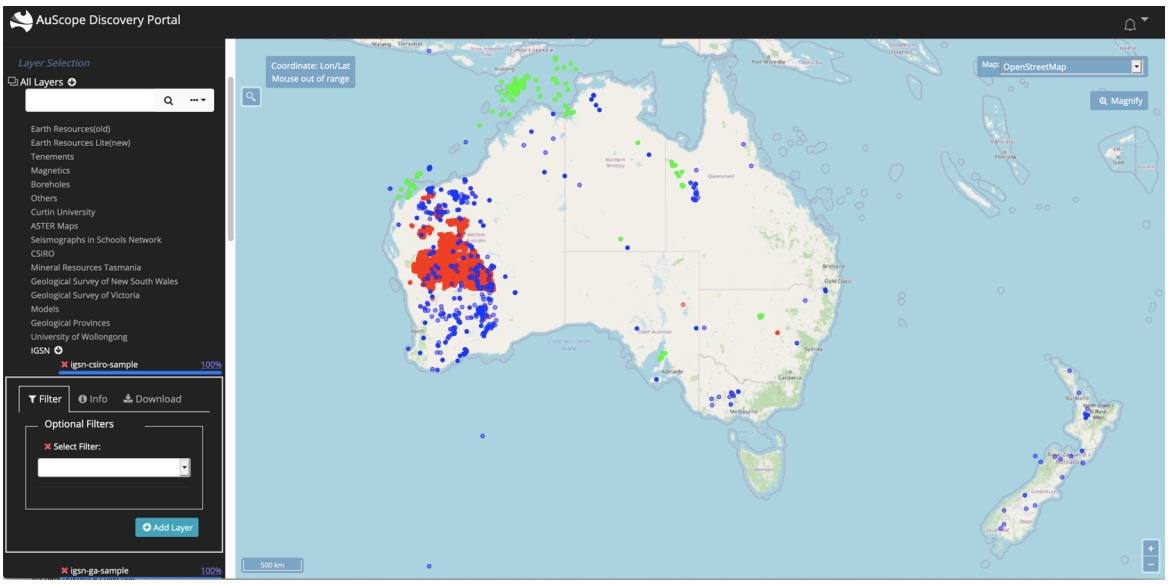


Figure 4. Map showing IGSN samples from CSIRO, ARDC, GFZ Potsdam and some samples from GA. However, the technology would not scale to include all 3.4M samples from Geoscience Australia.

GA has 3,414,664 samples registered with IGSNs but unfortunately the GeoDeVL project verified that the OAI-PMH is not capable of scaling to this number of samples. Tests on a new IGSN architecture being trialled by the International IGSN 2040 project based around JSON-LD have shown it will scale and enable maps to be created of the distribution of IGSNs in Australia, but this was not operational at the time of this project. Figure 4 shows the best effort achieved by the GeoDeVL project using the OAI-PMH.

[21] https://www.openarchives.org/pmh/

Output 2 of the IGSN work package aimed to investigate how to increase uptake of IGSN in Australian Geoscience Departments. In the first GeoDeVL project, ARDC developed an online service in which a user could get a single IGSN for a specimen, but quickly there was demand from at least four Australian Earth Science departments for a means of bulk minting: two departments had >10000 samples they wished to register. ARDC developed a bulk IGSN minter, but unfortunately only Curtin University succeeded in using it, whilst Melbourne University chose to mint IGSNs with the US System for Earth SAmple Registration (SESAR)[22]. The other interested universities did not have the infrastructure to support the ARDC IGSN bulk minter let alone the storage and linking to geochemical laboratory data and other data measured on these samples.

Work Package 4: AVRE

Most Australian geophysical datasets acquired with public funding are only accessible online as images: gridded forms of the data are available from websites as file downloads. In many cases, data are in proprietary formats, which limits uptake in the research community. In most cases Level 0, 1 and 2 forms of the data can only be accessed via direct contact to the collector/owner.

NCI developed its GSKY server to provide a new approach to online analysis and visualisation of datasets. GSKY provides an ability for users to interact with datasets, and any information they contain, using standard community protocols. GSKY[23] is specifically designed to access and analyse big geospatial datasets on NCI's cloud and HPC systems, and then deliver it to a user device or website. Furthermore, using GSKY's processing capability, the vast datasets can be easily analysed on the fly using user-provided algorithms to extract new information over both space and time.

The project initially proposed making the ASTER and the 2015/2016 National Geophysical Compilations datasets accessible through GSKY. The delay in the project was because GA released new versions of the National Geophysical Compilations in 2019-2020 during the lifetime of the project and it was felt that it was better to wait for the later versions. The ASTER and 2016/2019 National Geophysical Compilation GSKY WMS and WCS layers can now be

1. Discovered in NCI's GeoNetwork catalogue; and
2. Accessed as GSKY service WMS layers in the VGL and AuScope portals.

The methodology produced for publishing these National Coverages will be reused to publish web service endpoints for ~3000 individual radiometric, gravity and aeromagnetic surveys that are currently awaiting republication.

6. Achievements against activities & outputs/ milestones

Objective 1: Make 2 MT Time Series AusLAMP Surveys accessible in the NCI Catalogue and RDA

no.	activity	outputs/ milestones	completion date	comments

[22] https://www.geosamples.org/
[23] https://gsky.readthedocs.io/en/latest/

1.1	Form the AusLAMP Research Infrastructure Steering Committee, to be coordinated by AuScope	The group was formed in 2019 but has only met twice	September, 2019	The idea was that this group would help determine the standards, but no decisions have been made on the agreed standards for the collection and curation of Australian MT data. The GeoDeVL project was too far ahead of the decision being made and did the best it could.
1.2	Update the University of Adelaide Musgraves and University of Tasmania AusLAMP Datasets to be compliant with the new Australian Community metadata standards and if they are complete, the IRIS EMAC Technical Working Group for Magnetotelluric Data Handling and Software.	MT Data preparation for two time-series surveys	15-April-2020	Only one time series survey was completed, as GA requested the right to publish the Tasmanian AusLAMP dataset. Neither standard was finalised prior to beginning the HPC optimisation..
		MT HPC performance improvements (e.g., netCDF I/O)	30-June-2020	The project automated and optimised the processing and publishing of the Archived, Level 0 and Level 1 time series data from the Earth Data Logger instruments. Considerable speed performances were achieved: See Appendix 3.
1.3	Continue to ensure all new MT datasets added to the NCI Data platform as part of the GeoDeVL extension project are discoverable in the NCI catalogue, and then harvested into the RDA and AuScope catalogues/portals.	The different processing levels of the MT Musgraves time series dataset have been published on the NCI THREDDS Data server and are searchable through the NCI geonetwork catalogue. The datasets can be harvested by VGL AuScope Portal, VGL and RDA	15-April-2020	

Objective 2: To Report on potential for AusPass data to be accessible in Australian Research Data catalogues and portal

no.	activity	outputs/ milestones	completion date	comments
2.1	Determine how to make AusPass Passive Seismic data made accessible in RDA.	AusPass metadata consumed by the AuScope portal.	April 2020	The AusPass metadata data can be consumed by the AuScope portal, but cannot be connected to RDA, until 1) the AuScope portal is upgraded and 2) a translator is written from FDSN to ISO 19115.

Objective 3: IGSN Work Package: Investigating how to better support the Australian Research community in the Implementation of IGSN

no.	activity	outputs/ milestones	completion date	comments
3.1	In collaboration with AuScope Grid develop a searchable map of all IGSN samples	IGSN map is accessible in the AuScope portal	April 2020	A searchable map was developed, but because it was based on the OAI-PMH protocol, it would not scale to the number of samples minted by GA (3.4 M).

	registered in Australia			
3.2	Investigate the most sustainable way of machine to machine bulk uploading datasets from individual geoscientific laboratories, researchers and departments	The McNaughton collection at Curtin University successfully minted IGSNs using the ARDC bulk minting service	June 2020	Other universities were unable to achieve this as they did not have the support or supporting infrastructure.

Objective 4: To AVRE package

no.	activity	outputs/ milestones	completion date	comments
4.1	Update existing geophysics data to the new NCI OGC Compliant GSKY services	The CSIRO/GA/NTGS/GSSA/ AWAGS ASTER national datasets accessible via NCI's OGC GSKY services	31-July-2020	NCI geonetwork parent record: http://dx.doi.org/10.25914/5f224f36ec890
		Geoscience Australia 2015 versions of the National Geophysics (Magnetics, Radiometrics, Gravity) accessible via GSKY services	2-Oct-2020	We were able to publish the 2019 editions of Magnetics and Radiometrics as they became available during the project. The 2016 gravity was published. NCI geonetwork parent record: http://dx.doi.org/10.25914/5f76b265d7ce0
		Allow AVRE components to consume GSKY datasets	28-Sep-2020	
		Investigate potential of newer versions of these existing datasets as well as other data products available on the NCI data platform to be moved into GSKY data services.	2-Oct-2020	Added the latest (2019) versions of the National Geophysical Compilations Magnetic and Radiometric layers to GSKY.
4.2	Developing training notebooks to convert old VGL and other workflows into notebook technology	Enable notebooks to process data from GSKY using both WMS and WCS services.	2-Oct-2020	GSKY notebooks: ASTER WMS example ASTER WCS example National Geophysical Compilations WMS example National Geophysical Compilations WCS example

7. Key results and discussion

Work Package 1: Magnetotellurics

The project focussed on achieving a dirt-to-desktop-to-publication workflow (Figure 5). At the start of the project, data providers mainly made online processed MT EDI files and model outputs accessible as file downloads: the rawer time series datasets were only available through

direct request to the author/organisation that collected/owned the data. The data were then provided privately and often using physical media via the post. A lack of agreed community standards has meant that many processed EDI datasets from past surveys had inadequate metadata and 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 therefore reliant on the processing conducted by another MT scientist, which may or may not have met their target depth or processing requirements.

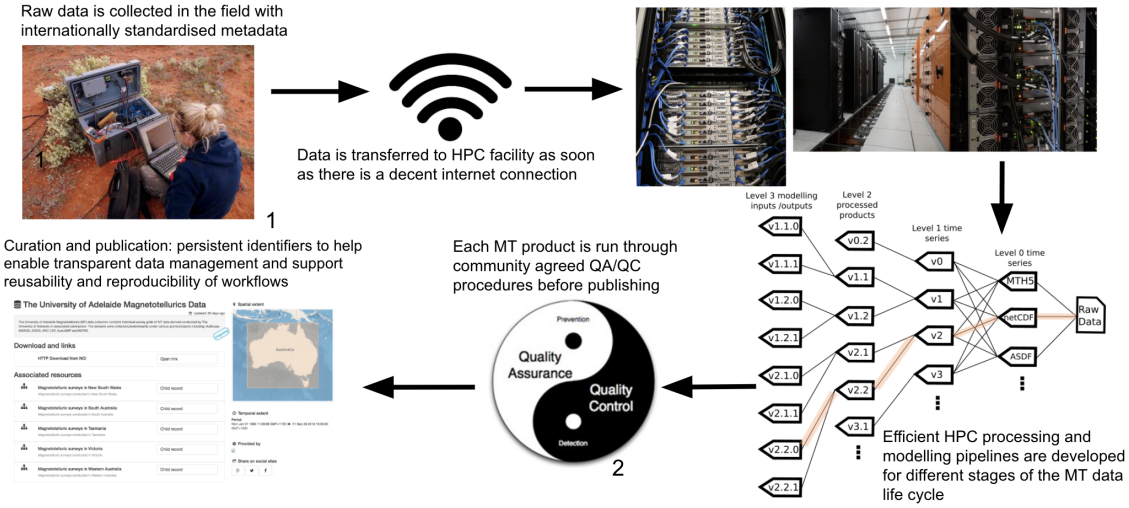


Figure 5: Dirt-to-desktop-to-publication workflow for magnetotelluric data/metadata.
1:http://energymining.sa.gov.au/minerals/geoscience/geological_survey/gssa_projects/auslamp
2:https://www.dialog.com.au/open-dialog/the-difference-between-quality-assurance-and-quality-control/

The MT time series datasets can be very large (100s of terabytes) and are unique: they can be very expensive to acquire or re-acquire, especially at continental scale. Because of the file sizes, most are stored in offline tapes, relatively inaccessible and potentially vulnerable to being lost. Hence, there is a growing demand for the rawer forms to be more accessible and secure than current practices allow, as they are required for replication, reanalysis and testing of new processing techniques. This project experimented with making calibrated MT time series datasets more accessible and aligned with the FAIR principles to increase reusability of the data and allow for more targeted processing to specific user needs. By translating the data into modern self-describing formats (i.e., netCDF) and then parallelising and porting the code to HPC and optimising it, the project showed that it was feasible for researchers that required it, to directly access the less processed forms of the data and generate their own EDI files that were more tuned to their specific use case. This is a substantial transformation of work practices.

The use of HPC was also critical, as to date, this step with the time series data had been undertaken on local processing infrastructures, mostly in serial, and was very time consuming. Accelerating time series publication codes via parallelisation (e.g., using MPI, OpenMP) combined with access to 1000s of CPUs allows a user to rapidly compute the different MT processing levels. Table 1 in Appendix 3 shows the summary of a benchmark test using the **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**MTtsdp**) codes on the Gadi supercomputer to process different time series processing levels for 95 Earth Data Logger stations (3328 different days of time series data). Table A3.1 in Appendix 3 demonstrates that a user can now process archived, level 0 and level 1 time series from a large MT survey in a matter of minutes.

We had no guidelines on best practice on how to make MT time series data accessible in online catalogues. In the interest of open, transparent science, the decision was made to publish the archived, Level 0 and Level 1 Musgrave time series datasets on the NCI THREDDS Data server (Figure 6) and make the associated metadata records searchable through the NCI geonetwork catalogue (Figure 7): data accessible so that a user can reprocess at any level.

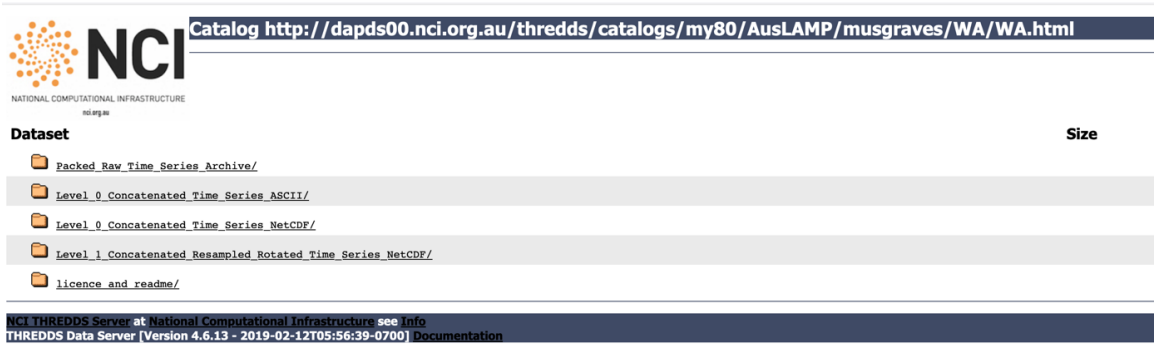


Figure 6: Musgraves AusLAMP time series data (Archived, Level 0 and Level 1) published on NCI's THREDDS data server

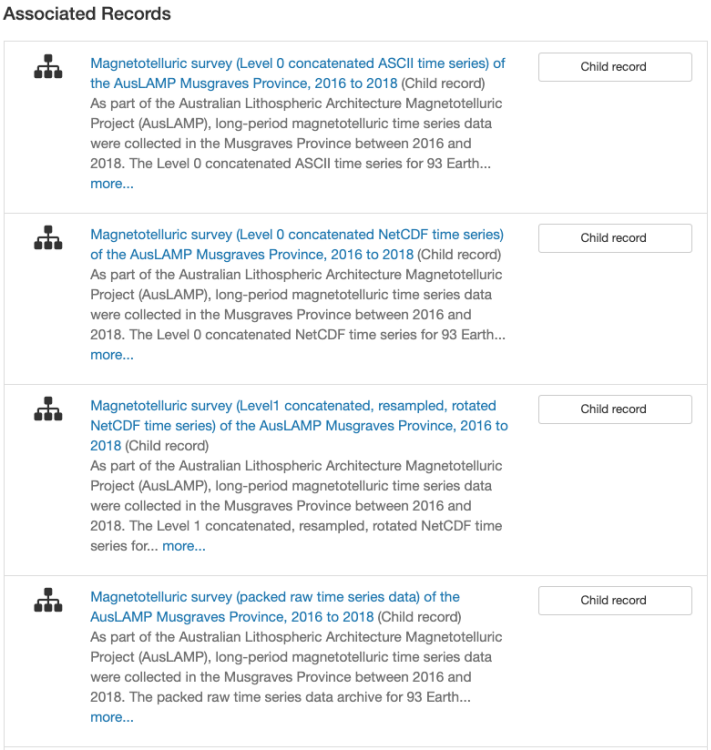


Figure 7: Musgraves Raw, Level 0 and Level 1 time series metadata records published on NCI's Geonetwork

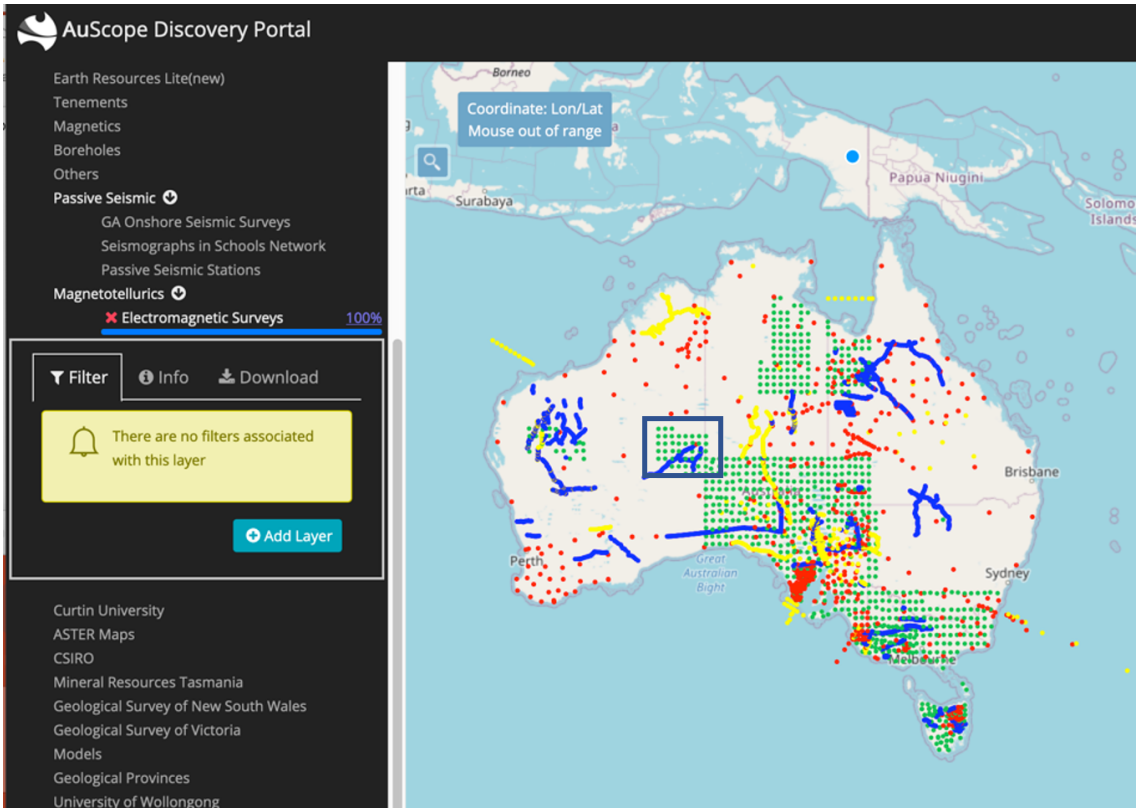
We have also developed the notebook tutorial *Accessing MT time series data from the NCI THREDDS Data Server using OPeNDAP*[24]. This tutorial demonstrates how a user can access the Level 1 AusLAMP time series netCDF files using OPeNDAP, query the metadata, visualise the

[24] https://nci-training.readthedocs.io/en/latest/_MT/MT_notebooks/THREDDS_OPeNDAP_crawler_my80_time_series.html

time series and process them in the frequency domain using the BIRRP code[25] to generate transfer functions, all without the need to download a single file. Processing time series in a Jupyter notebook would be far easier if the time series processing codes were available as Python libraries. IRIS have recently sent out an MT software development proposal [26] which includes the rewriting of Gary Egbert's MT time series data processing software (EMTF[27]) in Python.

By making the rawer MT time series products available online in self-describing formats that allow online web service access (e.g., OPeNDAP), a user can generate their own transfer functions using whatever codes they desire and transparently share their workflow via notebooks without the need to locally download large time series datasets.

However, although vast amounts of MT data have been collected in Australia with public funding in either government or research communities for decades (Figure 8), the Musgraves dataset is one of the few MT Time series dataset available online. It is also worth noting that much of the MT data in Figure 8 is also difficult to discover online as higher level data products (Level 2, 3), let alone access.



[25] https://www.whoi.edu/science/AOPE/people/achave/Site/Next1.html
[26] https://www.iris.edu/hq/files/mtdev/RFP_MT_Software.pdf
[27] http://www.mtnet.info/main/index.html

Figure 8: Electromagnetic surveys conducted in Australia with public funding. The lime green dots represent the long period AusLAMP array, yellow dots represent long period MT surveys, blue dots are broadband surveys, red dots are geomagnetic depth soundings. The black box shows the Musgraves dataset that the GeoDeVL project focussed on: time series data from all the other datasets are mainly available on request from the collectors/owners of the data.

The work done in the GeoDeVL project on time series data is 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 are now also planning to develop tools for specialized 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[28]. For some years, AuScope has been collaborating with international efforts in Earth Science Research Infrastructures (IRIS, UNAVCO, EarthCube, ESIP, EPOS and ENVRI) and it is essential to maintain these close linkages with those developing the global standards. In particular, in 2020, the NSF funded IRIS (seismology/MT data focus) and UNAVCO[29] (geodesy/GPS/GNSS data) groups are merging to form the EarthScope Consortium Incorporated and are already planning to develop a generalized container for geophysical time series, which will possibly be called GeoHDF and will be based on netCDF4. This would help overcome issues this project has faced with the debates over PH5, MTH5, ASDF and netCDF/HDF.

Work Package 2: Passive Seismic

The key result is that key Passive Seismic datasets are now off tapes and accessible online using the FDSN protocols that are used within the global passive seismic community. The GeoDeVL project focussed on making metadata on these data assets more widely accessible to the broader geoscience community through the AuScope Portal, but an updated crosswalk is required to make the metadata accessible in RDA.

Work Package 3: IGSN

At the start of the GeoDeVL project, it was realised that the technical design principles of the IGSN are more than ten years old and the shape of the IGSN system architecture had not changed since 2011: it was also based on XML. At the international level, it has been recognised that technology has moved on: dedicated servers have been replaced by cloud-based services and that XML as a way to structure information has been superseded by JSON and its dialects. Hence, in early 2020 the decision was made internationally to transition IGSN from XML-based metadata to web architecture based on sitemaps (e.g., schema.org) to enable extensible sample descriptions and accommodate many use cases. With this in mind, it was decided not to put too much effort into the existing IGSN infrastructures in Australia, and wait for the international community to move in unison to the new proposed infrastructure in 2021.

The main stumbling block in the uptake of IGSN in the research community is that the average Earth science department sample repositories and laboratories operate small scale support infrastructures, mostly based around spreadsheets on local hard drives or servers: they do not have the expertise or infrastructure for persistent database management. For example, one geology department has 33,698 registered sample numbers that are recorded in two spreadsheets, a set of index cards and a paper ledger : estimates are that they also have at least

[28] https://www.iris.edu/hq/rfp/mtdev.
[29] https://www.unavco.org/

10,000 unregistered samples. Curtin University succeeded because it had the institutional support of the Curtin Library. Other groups sought institutional support, but this was not available for physical samples.

To address this issue two proposals have been submitted to ARDC based partly on the findings of this GeoDeVL project:

1. **'SciDataMover:** Moving Long-tail lab and large-scale facility data FAIRly' was submitted to the ARDC 2020 Platforms proposal to address these issues, including the development of 'a repository of last resort' to try to address this dire situation.
2. **'AusGeochem**: An Australian Facility-Federated Geochemistry Data Repository' which seeks to develop a federated data repository from Australian Geoscience departments.

Work Package 4: AVRE

Currently the National Coverages for ASTER and geophysics datasets are mainly available from data catalogues as file downloads for local processing from multiple sites (e.g., CSIRO DAP[30], Digital Earth Australia[31], GA Catalogue[32], Figure 9). Where ASTER and national geophysics datasets are available on online sites, they are only available in GIS-style portals as either image files (WMS or GeoTIFFS - e.g., GA Portal[33], AusGIN Portal[34]) or files that can be downloaded. On some sites these image files have been subsampled to allow for faster loading of the images.

Given that the ASTER data is currently available from so many sites and in varying formats, one of the difficulties in publishing the ASTER datasets was related to understanding who had actually created the versions at NCI and who should be credited. Prior to publication, the project researched the complex lineage of the dataset. We also submitted it as a use case to the RDA Data Versioning Working Group to help align with best practice in citing datasets that have a complex provenance (Klump et al., 2020a, 2020b, in review).

[30] https://data.csiro.au/dap/landingpage?pid=csiro%3A6182
[31] https://data.dea.ga.gov.au/
[32] https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/search?facet.q=keyword%2FGA%20Publication&from=1&to=20
[33] https://portal.ga.gov.au/
[34] https://portal.geoscience.gov.au/

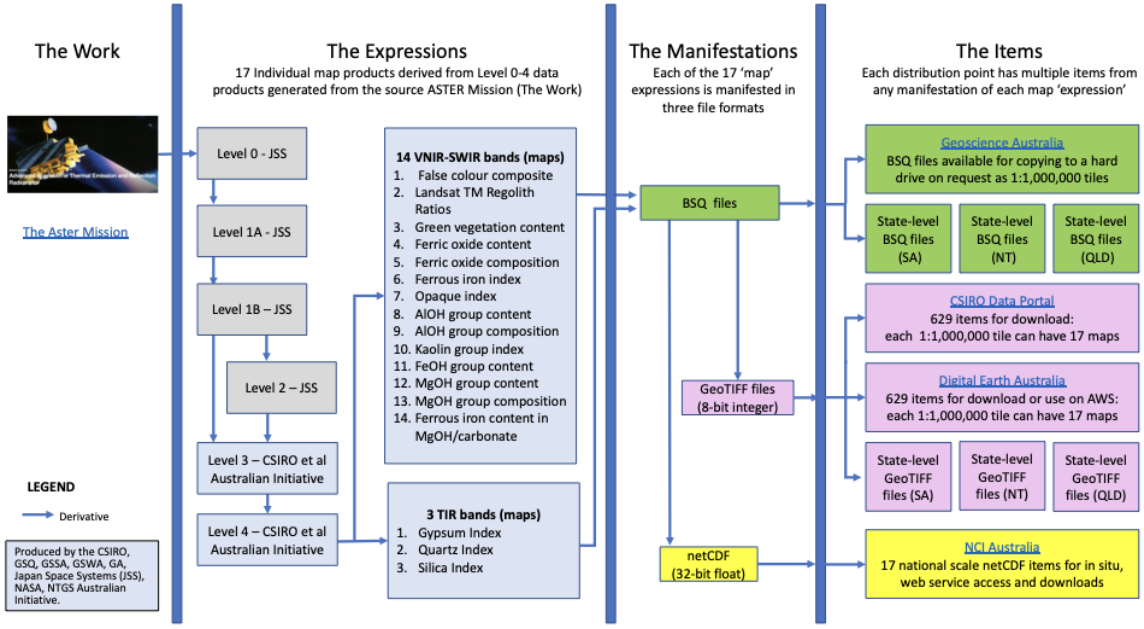


Figure 9. The complex provenance of the ASTER dataset at NCI and the different versions available from other sites.

The use of GSKY for the large National-scale ASTER and Geophysics coverages opens up new processing opportunities in high performance data analysis. GSKY enables the option for a user to both programmatically:

1. Visualise the ASTER and National Geophysical Compilation layers using the GSKY WMS endpoints; and
2. Access/download subsets of the rawer netCDF files using the GSKY WCS endpoints.

GSKY also has the capability of using the OGC Web Processing Service (WPS) which could be used in the future for defining request and response rules allowing for geospatial or geophysical processing on the fly.

The following four notebook examples were developed to demonstrate how a user could programmatically make GSKY WMS and WCS requests for both the ASTER and National Geophysical Compilation datasets:

1. ASTER WMS example
2. ASTER WCS example
3. National Geophysical Compilations WMS example
4. National Geophysical Compilations WCS example

With the launch of NCI's GSKY geospatial data server additional functionality was required by VGL to be able to search for, display and interrogate the new datasets. This required significant enhancements not just to VGL, but also in underlying support libraries that VGL relies upon in order to provide the necessary interfaces to NCI's Web Coverage Service (WCS) and Web Map Service (WMS) services that provide the data. Initially, this required adding the ability for VGL to query a single Catalog Service for the Web (CSW) record and then make an additional GetCapabilities request to the GSKY WMS service to retrieve a list of layers that GSKY provides. This most significant drawback of this approach was in data discoverability, as only having a single CSW record meant that for a layer to appear in VGL search results its relevant searchable

information (name, description etc.) must be present in the master CSW record. As the GSKY service provides access to many layers it meant the master CSW record would contain the details of many layers and thus be very difficult for a human to parse and find precisely what it is they are looking for. It was decided in the interest of best practices to instead provide an individual CSW record for every layer of the dataset, allowing VGL to more precisely search for the relevant information and display finer grained results to the user (Figure 10).

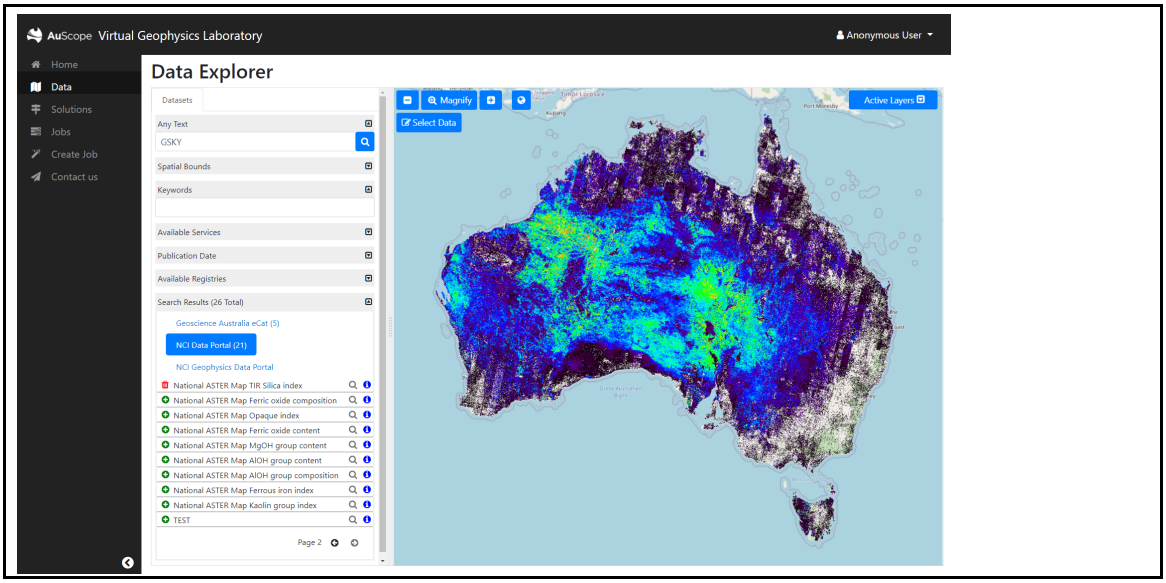


Figure 10: The VGL data portal showing GSKY search results and a layer displayed on the map.

On top of the GSKY WMS/WCS, NCI also provides OPeNDAP[35] and netCDF Subset Service[36] (NCSS) web services for the ASTER coverages via their THREDDS Data Server[37].

One of the key results of this project has been the ability for both the VGL and the AuScope portals to search and display results from both research and the government catalogues via a single interface (Figures 11, 12). In the AuScope portal, both the AusPass and the MT observing stations can be searched on together and information on the location and names of the sites examined (Figure 10). Dynamically connecting the data to each site will be a next step.

Figures 11 and 12 are proof of concept only. The location of the MT stations come from a database run by Graham Heinson which was converted into a test database at NCI[38] that could be accessed using the OGC Web Feature Service (WFS) protocol. The PS locations were provided as a file by AusPass. It would be desirable to develop a more permanent solution where both research and government organisations make the metadata of their station locations accessible for harvesting online, into an agreed Australian online database. It would be ideal if this

[35] https://www.opendap.org/

[36] https://www.unidata.ucar.edu/software/tds/current/reference/NetcdfSubsetServiceReference.html

[37] http://dapds00.nci.org.au/thredds/catalogs/wx7/catalog.html

[38] https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f7228_443_9506_3012

information became available soon after the data are collected: this could be developed in partnership with the ANSIR instrument pool.

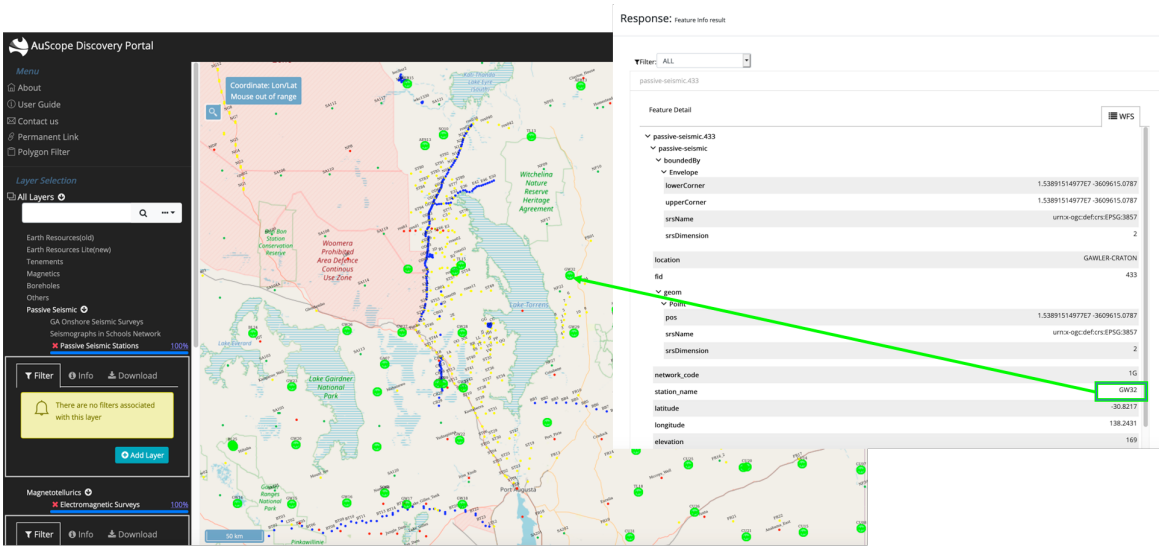


Figure 11. PS and MT stations in the one portal. Clicking on site GA32 produces the metadata for that site.

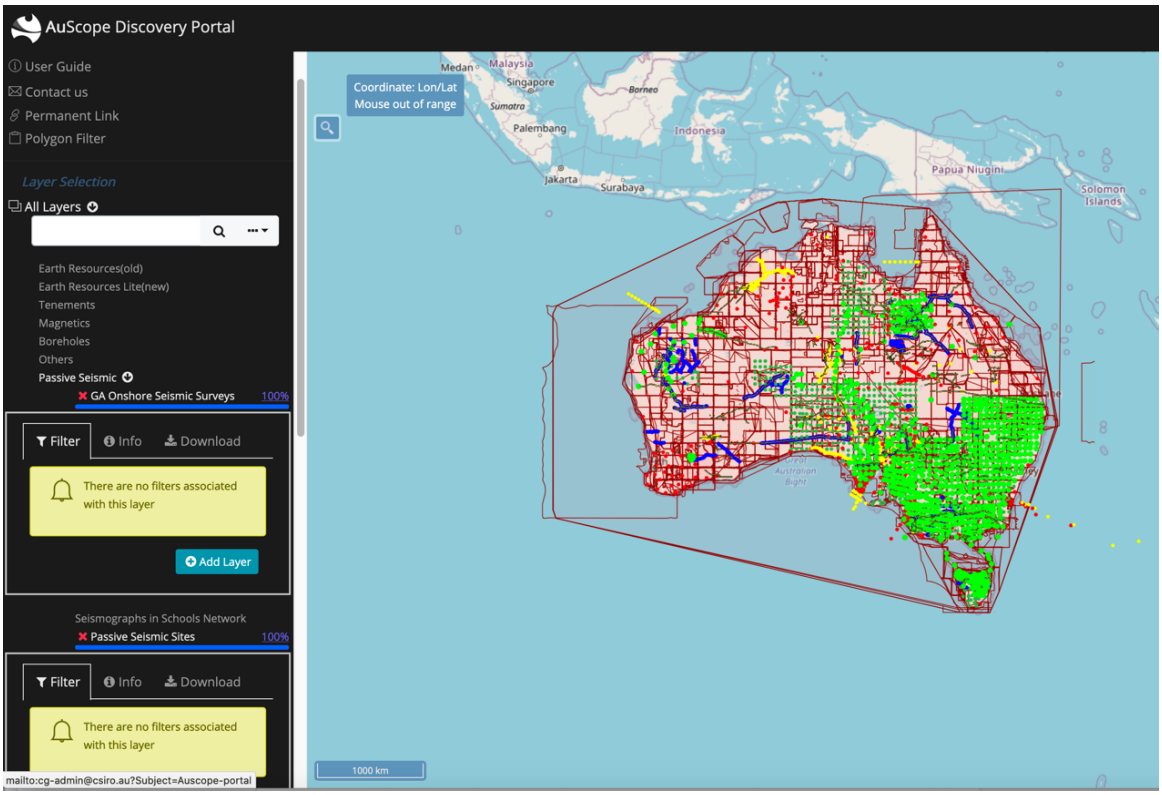


Figure 12. Location of GA seismic reflection surveys (black double lines) and Magnetic Surveys (brown polygons), PS sites from the ANU AusPass portal (lime green dots) and MT sites from Graham Heinson's MT surveys database (all other colours) are displayed making it easier for users to search an area and get the location of both government and academic research geophysical surveys of various techniques.

8. Project expenditure

Removed from public version of report

9. Impacts

For discussing impact we have used the OECD Reference Framework for Assessing the Scientific and Socio-economic Impact of Research Infrastructures[39] (OECD Science, Technology and Industry Policy Papers, March 2019, No. 65) and report against the relevant core impact factors listed in this document.

[39] https://www.oecd-ilibrary.org/docserver/3ffee43b-en.pdf?expires=1601284986&id=id&accname=guest&checksum=4CE284CBC72F1CFE6F71DDF641E88FE1

1. Scientific impacts – now and in 5 years

1. Open Science.

The development of the open and transparent processing pipeline, and in releasing all stages of the processing from Level 0 to Level 1, as well as Jupyter Notebooks on how this processing is done is in line with OECD Core Impact Factor - ***S12 Use and Production of Open Data***. This is also in compliance with Principle 5 of the Beijing Declaration on Research Data[40] which states that '*Publicly funded research data should be interoperable, and preferably without further manipulation or conversion, to facilitate their broad reuse in scientific research'*. It is hoped that within 5 years, all forms of publicly funded geophysical data from both the research and the government sectors will be more open and accessible than it is now.

2. Collaboration Excellence

During the lifetime of the GeoDeVL work packages, collaborations were formed with many equivalent Earth science research infrastructures internationally including IRIS, USGS, ESIP, EarthCube, EarthChem, ChEESE, EPOS. This is compatible with OECD Core Impact Factor of ***S9-Collaboration excellence (scientific).*** Particularly with the PS, IGSN and AVRE Work Packages being able to leverage existing tools, standards and infrastructures not only accelerated implementation, it raised awareness of what is being built in Australia and will help fast-track the development of research infrastructures that are globally interoperable. It is hoped that within 5 years, Australian geophysical data will be part of online global research infrastructure networks that will ultimately support the UN Sustainable Development Goals[41].

2. Capacity impacts – now and in 5 years

Work Package 1: Magnetotellurics

In this project we have prototyped new and open methods of making MT time series data FAIR and enhanced transparency. We have developed new HPC processing methods. One key impact is that were this HPC approach to be more widely adopted, the time taken to make MT 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 accessible for others to use.

Work Package 2: Passive Seismic

(Note: The AusPass portal was developed with funding in addition to the GeoDeVL). The combined funding has enabled an infrastructure to be developed whereby users can self-serve and obtain copies of the data. When the FDSN metadata can be cross-walked with other metadata standards it will create a new infrastructure where all Australian geophysical data from public funding can be discovered within a spatial area and accessed online.

Work Packages 1 and 2: Magnetotellurics and Passive Seismic.

The OECD core impact indicator ***S4: Number of Projects Granted***. Key findings from the MT and PS packages have led to two additional grants that will continue to grow the capacity developed as part of the GeoDeVL projects. These are:

1. **ARDC Platforms Program:** *Field Acquired Information Management System (FAIMS) 3.0:* In making PS and MT datasets FAIR, relevant field metadata was inconsistent and often

[40] https://www.codata.org/uploads/Beijing%20Declaration-19-11-07-FINAL.pdf
[41] https://sdgs.un.org/goals

stored in offline paper records. The FAIMS project is focused on digital loggers to ensure capture of standardised metadata attributes with field-deployed PS and MT geophysical instruments. A condition of using MT and PS instruments in the ANSIR pool will be that community agreed metadata attributes are collected and made accessible with the data.

2. **ARDC Cross-NCRIS Data Assets Program**: *Geophysics 2030: Building a National High-Resolution Geophysics Reference Collection for 2030 Computation* was based on the findings of the MT and PS Work Packages and seeks to use the MT and PS datasets as pathfinders on how to move other minimally processed, high-resolution geophysics datasets to the NCI collections to create an integrated national high-resolution reference collection of other geophysical types (magnetic, gravity, AEM, radiometrics, INSAR, Distributed Acoustic Sensing (DAS), GRACE, etc) suitable for HPC data analysis.

Work Package 3: IGSN

Any impacts for the ARDC IGSN bulk minting service will not be realised until more departments have access to infrastructure that enables them to maintain and store databases. To build capacity, three proposals have now been submitted:

1. **ARDC 2020 Platform Program:** *SciDataMover: Moving Long-tail lab and large-scale facility data FAIRly.* This seeks to develop a discipline- and scale-agnostic, lightweight, open source Data Movement Platform that transfers coupled data/metadata from laboratories to shared workspaces then to an existing repository. The proposal does include a 'Repository of Last Resort' for those departments that cannot get domain or institutional support for IGSN allocation and/or storage of their data.
2. **ARDC Data Partnerships Program:** *AusGeochem - An Australian Facility-Federated Geochemistry Data Repository.* The aim is to aggregate a globally significant volume of Australian publicly funded geochemical and geochronological data, and will leverage the ARDC IGSN minter to ensure globally unique identification of samples.
3. **ARDC 2019 Platform Program: *FAIMS 3.0***: *Electronic Field Notebooks*[42]. In the context of modernising the FAIMS electronic notebook app framework, modules for documentation of samples at the point of origin and linking to IGSN will become available.

Work Package 4: AVRE

The focus was switching from rigid work flows to more flexible systems, whilst at the same time lowering the technical debt incurred from maintaining little used components. A collection of Jupyter Notebooks have been produced to enable this. The requirement for VGL to be able to parse GSKY datasets has also resulted in modifications to VGL and underlying libraries that allow it to search for, display and interrogate geophysical datasets that it previously was unable to, that is, layers that do not possess individual CSW records and are only defined via a GetCapabilities request to the server's WMS service. This functionality allows VGL to access a broader pool of datasets and provide users with even more datasets to discover and analyse. The VGL portal will ultimately be merged with the AuScope portal.

3. Community impacts – now and in 5 years

[42] https://ardc.edu.au/project/faims-3-0-electronic-field-notebooks/

One of the key OECD core impact factors is ***S10 Structuring effects on the RI on the Scientific Community.*** The MT Work Package has effectively restructured the way rawer forms of MT data are accessed and processed. Not only is the rawer data now more accessible to the research community, tools have been developed to dramatically reduce the time taken for this processing. In effect, the project was building infrastructure that is creating cultural change. For example, although there was a small community that was advocating for greater access to, and transparency in processing of the rawer forms of MT data, in some groups, both Nationally and Internationally this is not seen as necessary, and these groups still advocate for only making the highly processed data products accessible. The enabler of change has been the greater accessibility of high performance data and compute infrastructures such as those at NCI.

With the IGSN work package, there is a growing community in Australia that want to use IGSNs but the limitations on the infrastructures in their local departments mean that they cannot easily utilise online tools such as the ARDC IGSN bulk minter.

The NCI catalogue and the AVRE/VGL portals enable in situ access to rawer forms of the data and online linkage with HPC facilities. In contrast, most online 'data portals' only provide access to images of the data, and GIS 'data layering' functionality, or functionality that enables data selections to be 'clipped', 'shipped' and downloaded for local processing.

4. Economic impacts

The OECD lists many core impact factors that relate to uptake by industry of research including ***T20: Innovations co-developed with industry; T21: Joint technology development projects between RI and industry T23 Projects funded by companies; T24 Collaborative projects with industrial partners; T27 Data Sharing and T28 Data commercial use and data services.*** The minerals exploration industry has made substantial investments in collecting MT data and hence the key question is whether the developments of the GeoDeVL are transferable to this industry?

Although the Musgraves dataset sampled LP MT data, which is not a direct input for prospect- and district- based minerals exploration as it is too deep in the crust, it has commonalities with many other geophysical time series data that sample different physical properties (Figure 13). All of these data types have the same issue in that the rawer forms of the data are not easily accessible and publishing techniques developed by this project would be transferable to them.

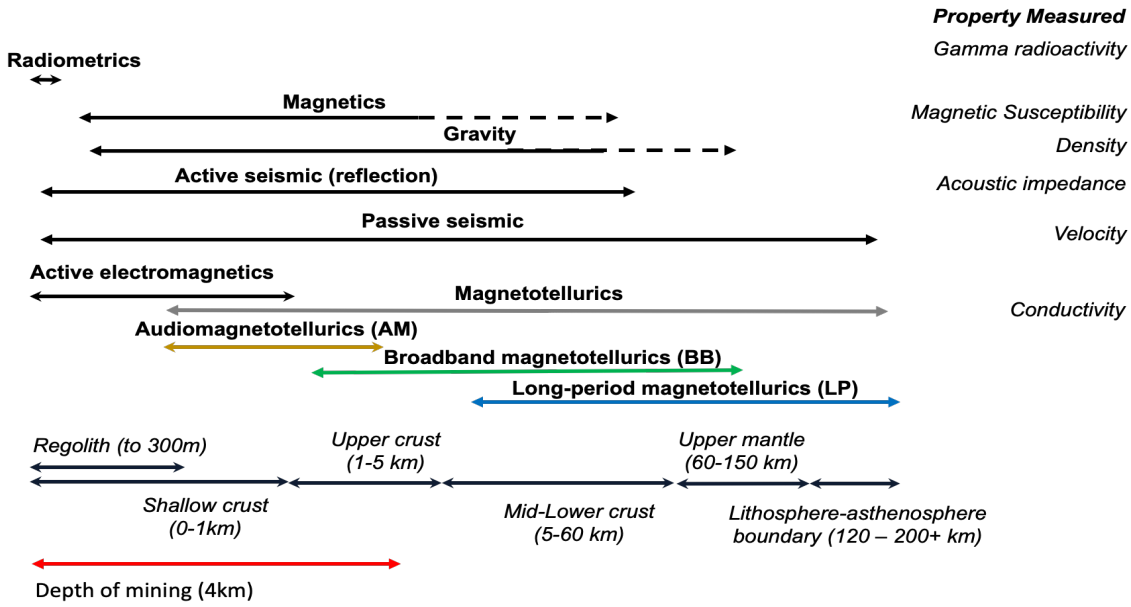


Figure 13. Types of geophysical data collected in Australia, the physical property measured and the depth of the crust that it sampled: also shown is the depth of current mining. Figure modified from Richard Chopping.

HPC computation is not widely used by the minerals exploration industry in Australia, unlike the petroleum industry that has embraced HPC for processing of geophysics data over the past two decades. At least 12 of the Top 500 supercomputers in June 2020[43] are focused on oil industry applications, but the minerals industry has been slow to adopt HPC techniques. In part it is because the minerals industry in Australia is dominated by small to medium enterprises that do not have access to HPC - many 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 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, these barriers in the minerals industry can be broken down and that bridges can be built between the research and minerals exploration industry to trial HPC, particularly for effective prospect/district scale targeting.

The key economic impact of this project is that the speed up in processing times means that the turn around for results is way faster, which was one of the key drivers for uptake of HPC in the petroleum industry.

5. Social impacts

The social impacts are yet to be realized but easier access to HPC and high volume data infrastructures is likely to create 'societal' impacts from this work. For example any researcher can now access the rawer, full resolution L0, L1 levels of the data and this can create nervousness amongst the data collectors/providers who in the past have been the only ones that can produce the higher level L2, L3 data products: hence some are reluctant to release the

[43] https://www.top500.org/lists/top500/2020/06/

rawer data. Some are also reluctant to share processing methodologies as these are seen as competitive advantages. However, it has to be noted that in the past, even if data collectors/providers wanted to share the rawer forms of the data, infrastructure limitations prevented them from doing so.

6. Environmental impacts

The results are too preliminary to gauge potential environmental impacts.

10. Communication and dissemination activities

Note: COVID-19 has significantly impacted on dissemination activities - particularly conferences.

All Work Packages:

Wyborn, L., Evans, B., and Robinson, E., 2019. What must the 'Big Data-HPC Ecosystem' look like in 2030 for Australian Geoscience research to be internationally competitive? Australian Leadership in Computing Symposium, Canberra, November 2019, http://dx.doi.org/10.13140/RG.2.2.26339.30244

Wyborn, L., 2020. Towards World-class Earth and Environmental Science Research in 2030: Will Today's Practices in Data Repositories Get Us There?, EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-22478, https://doi.org/10.5194/egusphere-egu2020-22478

Wyborn, L., Evans, B., Rawling, T., Fraser, R., Whiteway, T., Burton, A., 2020a. Competitive Solid Earth Science Data-Intensive Computational Research in 2030: What Will it Look Like? C3DIS 2020 conference, Melbourne, 2020. http://www.c3dis.com/3856

Wyborn, L., Rees, N., Woodman, S., Friedrich, C., Rawling, T., Heinson, G., Klump, J., Martin, J., Benn, J., Salmon, M., Evans, B., Fraser, R., 2020b. The Last Dance of the GeoDeVL: Actions Arising to Further Grow the Capacity of the AuScope Downward Looking Telescope. C3DIS 2020 conference, Melbourne, 2020b. http://www.c3dis.com/3862

Work Package 1: Magnetotellurics

Druken, K., Evans, B., Gohar, K., Rees, N., Richards, C., Wang, J. and Yang, R., 2019. Improving data reusability for high performance dataset. Abstracts 2019 AGU Fall Meeting, San Francisco, USA. https://ui.adsabs.harvard.edu/abs/2019AGUFMIN21A..10D/abstract

Rees, N., Conway, D., Yang, R., Evans, B. and Heinson, G., 2019a. Reusing Historic Magnetotelluric Time Series Data on HPC Infrastructure. Abstracts 2019 AGU Fall Meeting, San Francisco, USA. https://ui.adsabs.harvard.edu/abs/2019AGUFMIN23D0906R/abstract

Rees, N., Heinson, G., Evans, B., Rawling, T. and Wyborn, L., 2019b. Facilitating Future Reuse of Magnetotelluric (MT) Data: Moving to an Online Dirt-to-Desktop (D2D) Data Path to Standardise Data Collection, Curation and Publication. Abstracts 2019 AGU Fall Meeting, San Francisco, USA. https://ui.adsabs.harvard.edu/abs/2019AGUFMIN24A..08R/abstract

Work Package 3: IGSN

Lehnert, K., Wyborn, L., Klump, J., Elger, K., Carter, M., and Fleischer, D., 2019. Digital Infrastructure and Policies to Support Open and FAIR Physical Samples and Collections in the Earth Sciences. CODATA 2019 Conference on Next-Generation Data-Driven Science, Beijing, China, September 2019. https://conference.codata.org/CODATA_2019/sessions/88/paper/525/

Klump, J., Lehnert, K., Wyborn, L., and Ramdeen, S. and the IGSN 2040 Steering Committee, 2020. Building a sustainable international research data infrastructure - Lessons learnt in the IGSN 2040 project, EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-12001, https://doi.org/10.5194/egusphere-egu2020-12001

Work Package 4: AVRE

Fraser, R., Rawling, T., Klump, J., Woodman, S., Friedrich, C., Warren, P., Fazio, V., Wyborn, L., 2020. The AuScope Data Portal: supporting adaptable, next generation data-intensive convergent research. C3DIS conference, Melbourne, Australia, March 2020. http://www.c3dis.com/3860

11. Conclusions and recommendations

A. Conclusions

Work Package 1: Magnetotellurics

Once community standards have been established, dirt-to-desktop-to-publishing pipelines can be developed to rapidly generate the different processing levels of MT data in a computationally reproducible manner. We have demonstrated what this might look like with the Musgraves MT time series data.

Work Package 2: Passive Seismic

During the lifetime of the GeoDeVL project, and with funding from other sources, the AusPass Portal has made numerous datasets accessible online.

Work Package 3: IGSN

ARDC have developed a valuable IGSN minting service to the Research community and there is a considerable desire to use this. However, at the departmental level, there is just not enough infrastructure capacity and capability to deploy it. Without institutional support for most departments, a facility-federated repository is rapidly becoming the only option. The current issues in using the AuScope portal to display Australian IGSN located samples will be resolved once the new IGSN architecture is deployed.

Work Package 4: AVRE

With the implementation of the GSKY service it was necessary to adapt both the AuScope and VGL portals to provide users with access to Aster datasets. Both portals were successfully adapted to be able to communicate with the service and search it for Aster layers, display those layers, and download requested areas of interest in formats necessary to run VGL jobs. The work in this project has shown that it is possible to use data services and make geophysics datasets available for online processing.

Further, it has laid the foundations for the 2030 Geophysics project in which multiple, large volume, high-resolution geophysical datasets, as listed in Figure 12, can be accessible on a single platform for advanced computation that combines analytical methods such as inversions and simulations, as well as real-world observations to develop probabilistic descriptions of specific Earth processes at any depth in the crust and upper mantle.

B. Recommendations

Work Package 1: Magnetotellurics

- For MT data to be Findable, Accessible, Interpretable and Reusable (FAIR) at scale, it needs to be both human and machine actionable, 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.

- Once the required (meta)data standards have community agreement, efforts should be made to have the standard peer-reviewed and endorsed by the International Association of Geomagnetism and Aeronomy (IAGA) of the International Union of Geodesy and Geophysics. The US community is planning to do this with the IRIS/EMAC standard.

- Once the standard is endorsed and stable, approaches should be made to the instrument companies to make their measurements and outputs compliant with the standard. This approach is similar to how seismology instruments adopted the SEED standard.

- Digital Field Data Loggers are seen as essential to developing an efficient Dirt-to-Desktop processing system to ensure that all MT and PS field data are born digital, FAIR, and born connected to a trusted digital repository. In rescuing data in both the MT and PS Work Packages, finding the critical field acquisition metadata attributes proved to be the hardest part. The new ARDC-funded FAIMS project will work towards addressing this.

- Currently obtaining information on the actual location of many Australian PS and MT survey sites collected with public funding is not easy: the information is either in private databases or only available as location files within published datasets. In many cases, the locations are not revealed until the research publication based on the sites is released. From the work of the FAIMS project, combined with the ANSIR Instrument pool and with collaboration of the government agencies it should be feasible to aim for a publicly accessible database that harvests station locations from distributed sources.

- DOIs should also be assigned to each station, to enable tracing of where data from a station is used in a publication. Currently, if DOIs are assigned, it is only at the network/survey level.

- Large national scale MT surveys that are collected over a decade (such as AusLAMP) could greatly benefit from HPC by making the different MT data processing levels available in consistent high performance data formats in a community project on a HPC system such as Gadi. This way, the whole MT community could have access to both raw and further processed AusLAMP data: they could easily update earlier surveys to input parameters and processing algorithms of their choice.

- As MT surveys get larger and are collected at higher resolution and closer spacing, there is a need to further develop highly parallelised HPC codes that rapidly and transparently produce the different MT processing levels.

- The international standards for MT are changing rapidly, and in line with the AuScope 10 year Strategy for 2020-2030 geoscience of applying global data principles and data management best practice, close contact needs to be kept with those groups that are developing FAIR MT standards and equivalent research infrastructures for MT data.

Work Package 2: Passive Seismic

- The GFZ Potsdam repository has a translator of FDSN metadata to the DatCite metadata profile. Given that the ARDC Data Retention Program[44] is using a profile of the DataCite metadata, it is feasible this could be leveraged to get the AusPASS data into RDA.

Work Package 3: IGSN

- The underlying issues of deploying IGSN effectively at the research department/ institution level cannot easily be resolved. They are not unique to IGSN identifiers, but are common for many Long Tail datasets collected by small numbers of researchers working in departmental/institutional laboratories. There needs to be a review on how to better support IGSN, not just in Earth science departments, but in any academic department/institution wishing to use IGSN for uniquely identifying physical samples.

Work Package 4: AVRE

- This project has further refined work done by the AuScope and VGL projects in making rawer forms of data accessible online. Unfortunately in Australia there are a plethora of portals in the government and research sectors, and although each services specific business needs/clients/stakeholders/etc., it makes discovering and accessing information across all of them very difficult, particularly the government and academic sectors. The solid Earth science community needs to work together for a more coherent approach to the access and delivery of datasets, particularly the rawer L0 to L2 levels. It could look to mimicking for example the Australian Oceans Data Network[45], which provides a high-level one-stop-shop for the simple discovery of, and access to, marine and climate datasets from universities, government, non-government and industry sites.

12. References

1. References cited in the report

Kirkby, A., 2019. Developing metadata standards for time series magnetotelluric data. Preview, 2019:199, 49-53, DOI: 10.1080/14432471.2019.1600210. [Last accessed 9 October, 2020].

Rees, N., Evans, B., Heinson, G., Conway, D., Yang, R., Theil, 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

[44] https://ardc.edu.au/our-strategy/storage-and-compute/
[45] https://portal.aodn.org.au/

Performance Computing Platform. ASEG Extended Abstracts, 2019:1, 1-6, DOI: 10.1080/22020586.2019.12073015. [Last accessed 9 October, 2020].

Wilkinson, M.D, Dumontier, M., Aalbersberg,I .J., Appleton, G., Axton, M., Baak, A. ... & Mons, B., 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18. [Last accessed 9 October, 2020].

2. List of publications produced by project

Klump, J., Wyborn, L.A.I., Downs, R.R., Asmi, A., Wu, M., Ryder, G., and Martin, J., 2020a. Compilation of Data Versioning Use cases from the RDA Data Versioning Working Group. Research Data Alliance. Available at https://doi.org/10.15497/RDA00041 [Last accessed 9 October, 2020].

Klump, J., Wyborn, L.A.I., Wu, M., Downs, R.R., Asmi, A., Ryder, G., and Martin, J., 2020b .Final Report of the Research Data Alliance Data Versioning Working Group - Principles and best practices in data versioning for all datasets big and small. Research Data Alliance. Available at https://doi.org/10.15497/RDA00042 [Last accessed 9 October, 2020].

Klump, J., Wyborn, L.A.I., Wu, M., Martin, J., Downs, R.R., and Asmi, A., in review. Principles and best practices in data versioning for all datasets big and small. Submitted to Data Science Journal.

13. Appendices

Appendix 1: Impact Stories

Impact Stories

Appendix 2:

A review of standards used to describe MT data

In late 2019, prior to the preparation and publication of the MT time series in the GeoDeVL project, a review was undertaken of standards being used/being developed/being proposed by both the International and National MT Communities in the research and government sectors.

The following groups of MT data standards and formats were then considered by the project:

1. **Metadata standards for time series data:**

For metadata standards there were two options available:

a) The Australian MT community published a metadata standard in May 2019 in the Australian Society of Exploration Geophysicists (ASEG) Preview paper on 'Developing metadata standards for time series magnetotelluric data'[46]; and

b) The Standard for Exchangeable Magnetotelluric Metadata was published by the Technical Working Group for Magnetotelluric Data Handling and Software[47] of the US NSF-funded Incorporated Research Institutions for Seismology (IRIS)[48] ElectroMagnetic Advisory Committee (EMAC)[49] in July 2020. It is also available in JSON-LD[50].

2. **Variants of HDF5/netCDF4**

There are at least four versions of the HDF5/netCDF4 being considered/used including:

a) The Adaptable Seismic Data Format (ASDF)[51] which has been used in Australia for MT data (e.g., Duan, J., and Kirkby, A., 2018. Data Formats, interoperability and sharing. Presentation to EMIW Workshop, Helsingør, Denmark, 2018).

b) The Portable Array Seismic Studies of the Continental Lithosphere (PASSCAL) Hierarchical Data Format Version 5 (PH5)[52] being used by IRIS for archiving MT data. However, IRIS use MTH5[53] for exchange of data with the other groups using MTH5;

c) MTH5 is used by the US Geological Survey (USGS) for archiving, exchange and processing; and

d) Plain netCDF4 was used by NCI for the Musgraves AusLAMP dataset in the GeoDeVL project (mainly because a, b, and c were still in test/prototype).

With the recent merger of IRIS and UNAVCO[54] to form the EarthScope Consortium Incorporated, plans are now being made to develop a generalized container for geophysical time series data

[46] https://www.tandfonline.com/doi/full/10.1080/14432471.2019.1600210
[47] https://www.iris.edu/hq/about_iris/governance/mt_soft
[48] https://www.iris.edu/hq/
[49] https://www.iris.edu/hq/about_iris/governance/emac
[50] https://github.com/kujaku11/MTarchive/blob/tables/docs/mt_metadata_guide.pdf
[51] https://seismic-data.org/)
[52] https://www.passcal.nmt.edu/content/ph5-what-it
[53] https://github.com/kujaku11/MTarchive
[54] https://www.unavco.org/

based on NetCDF, which will possibly be called GeoHDF (note: the name is not definite)). The aim is to replace PH5 and broaden the scope to include Distributed Acoustic Sensing (DAS) seismic data, MT, etc. The this generalized container for geophysical time series (and if possible other types) aims to hold data for IRIS, plus UNAVCO, plus work for field use with sensors from the PASSCAL instrument pool[55] (which is equivalent to ANSIR[56] research Facilities for Earth Sounding). Scoping work is commencing on this project now.

3. **Other publicly available standards used for MT metadata/data:**
   a) The USGS developed EMTF XML (EMTF XMF: New data interchange format and conversion tools for electromagnetic transfer functions[57]);
   b) GFZ Potsdam have published 'EMERALD a Data Format for Magnetotelluric data'. EMERALD uses Extracted Files (XTR/XTRX) which are ASCII files which are internally structured using the Extensible Markup Language (XML)[58];
   c) The Society of Exploration Geophysicists MT/EMAP Data Interchange Standard (1987) (i.e., the Electrical Data Interchange (EDI) file)[59]; and
   d) Geological Survey of Queensland's publicly available vocabularies that go with their new Geoscience Database[60].

[55] https://www.passcal.nmt.edu/
[56] http://ansir.org.au/
[57] https://library.seg.org/doi/10.1190/geo2018-0679.1
[58] https://gfzpublic.gfz-potsdam.de/rest/items/item_1284934_11/component/file_1284980/content
[59] https://www.seg.org/Portals/0/SEG/News%20and%20Resources/Technical%20Standards/seg_mt_emap_1987.pdf
[60] https://vocabs.gsq.digital/vocabulary/

Appendix 3: Comparative Times for Processing the MT Time Series Data.

Step	***What***	***No of Sites Processed/ CPUs used on GADI***	***New time to complete work on GADI***	***Estimated time using parallelised code on average machine with 4 cores***	***Estimated time for "site-by-site" processing (i.e., without automation) on average machine with 4 cores***
***Packed Raw Time Series Archive* (pack.py)**	Creates a zip file of the raw instrument data for each site in the survey (e.g. station1.zip, station2.zip,....).	95 sites processed using 96 CPUs	17 minutes	6-7 hours	Multiple days
***Level 0 Concatenated Time Series ASCII* (time_series.py)**	Creates Level 0 concatenated EX, EY, BX, BY, BZ ASCII files per station per day. No associated metadata available.	95 sites processed (3328 different days) using 96 CPUs	3 minutes	1-2 hours	Multiple days
***Level 0 Concatenated Time Series netCDF* (netcdf.py)** →	Creates a single Level 0 concatenated netCDF file per station per day for variables EX, EY, BX, BY, BZ. Additionally, the MT time series metadata is attached directly to the netCDF file.	95 sites Processes (3328 different days) using 96 CPUs	2 minutes	1-2 hours	Multiple days to weeks
Level 1 Concatenated Resampled Rotated Time Series netCDF **(01_check.py, 02_merge.py, 03_ascii_2_bin.py, 04_rotate.py, 05_make_netCDF.py)** → → to complete job.	Checks raw time series for gaps and if gaps exist, those problematic stations can be segregated for stage 2 processing. The stations that pass the checks are merged into concatenated (over **ALL** days) ASCII files (EX, EY, BX, BY, BZ). These ASCII files are then converted into a temporary binary file to accelerate the subsequent I/O operations. The binary data are read in, downsampled and rotated to north. The outputs are converted into a netCDF file (single file per station) and all the metadata attributes are added to the header.	Processed 83 sites using 16 CPUs	8 minutes	1-2 hours	Multiple days to weeks

Table A3.1: Benchmark test for processing different MT time series processing levels using the **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**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. This presentation[61] here provides more details on the processing methods used.

[61] https://docs.google.com/presentation/d/1MJfp20AlTBID9WbuU8UJXNHDdvXvKHwv53aWRA1OuS8/edit#slide=id.g817d8b9c14_0_52