Give the GeoDeVL her dues: she has helped lay the foundations for a modern eResearch platform for Solid Earth Scientists

Lesley Wyborn[1], Nigel Rees[1], Carsten Friedrich[2], Tim Rawling[3], Ben Evans[1], Graham Heinson[4], Michelle Salmon[5], Meghan Miller[5], Julia Martin[6], Joel Benn[6], Geoff Squire[2], Stuart Woodman[2], MingFang Wu[6], Ryan Fraser[7], Adrian Burton[6]

[1]National Computational Infrastructure, ANU, Canberra, Australia, Firstname.Lastname@anu.edu.au
[2]Data 61, CSIRO, Australia, Firstname.Lastname@data61.csiro.au
[3]AuScope Ltd, Melbourne Australia, tim.rawling@unimelb.edu.au
[4]The University of Adelaide, Adelaide, Australia, graham.heinson@adelaide.edu.au
[5]Research School of Earth Sciences, ANU, Canberra, Australia, Firstname.Lastname@anu.edu.au
[6]Australian Research Data Commons, Australia, Firstname.Lastname@ardc.edu.au
7Mineral Resource, CSIRO, Perth, Australia, ryan.fraser@csiro.au

SUMMARY

Since 2007, AuScope has been the national provider of research infrastructure to the earth and geospatial sciences communities in Australia and has supported a diverse suite of data generating infrastructures ranging from VLBI telescopes to geochemistry/geochronology laboratories and geophysical data acquisitions. In parallel, AuScope supported the development of data access portals, a suite of numerical simulation and inversion codes, and built virtual laboratories and online processing systems that connect these. New directions outlined in the 2016 National Research Infrastructure Roadmap called for the development of an integrated Downward Looking Telescope which requires distributed observational, characterisation, and computational infrastructure to provide the capability for Australian researchers to image and understand the composition and characteristics of the Australian Plate with unprecedented scale and fidelity. The GeoDeVL projects have laid the foundations for the transition to a modern platform for solid Earth researchers to support the Downward Looking Telescope. The new systems also need greater alignment of samples, data and software with the FAIR principles: in particular they must be machine actionable [1].

INTRODUCTION

From 2017 to 2019 the Australian Research Data Commons (ARDC) has co-funded three GeoDeVL projects in collaboration with AuScope, National Computational Infrastructure (NCI), CSIRO, The University of Adelaide, Geological Survey of South Australia, Research School of Earth Sciences (RSES) of the Australian National University (ANU), and Curtin University. The GeoDeVL projects were seen as a first step in realising the AuScope Virtual Research Environments (AVRE), a data assimilation and geoscientific discovery and analytics platform for the Australian continent to support the Downward Looking Telescope. AVRE is built around three related networks: Geophysics, Geochemistry, and Geology.

The GeoDeVL projects had their origins in the Virtual Geophysics Laboratory (VGL), which 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. There were many other digital geoinformatics and computational geoscience projects that were subsequently funded by groups such as AuScope, ANDS, NeCTAR, RDS, GA, CSIRO, the State and Territory Geological Surveys, ARC, Universities, etc. But despite best intentions, by 2015 it was realised that the systems developed in these projects were disjointed and tended to reflect the sector they came from and/or the capabilities of those being targeted (i.e., research, government, industry, public, etc.).

A review by AuScope in 2015 of the state of play in geoinformatics and computational geoscience in Australia noted:

1. There was no single portal to discover/access geoscience data from research, government and industry: rather there were a plethora of portals each with their own User Interfaces (UI), formats and tools.
2. There were too many virtual laboratories, which were very confusing to the user, each had different UIs and workflows, and overall were too expensive to maintain and continually update;
3. Academic Geophysics Data Collections, particularly for Magnetotellurics (MT) and Passive Seismic (PS) were not discoverable or accessible online: most were only available through 'contacting the author';
4. Academic geochemistry and geochronology data was also mostly inaccessible and there was no canonical way of finding the source samples and connecting them to data and publications derived from them; and
5. Many software tools to process, simulate, and model geoscience data had been developed, but they could not be easily discovered or accessed, and there was much reinventing of the wheel.

As AuScope's 2016 vision of a Downward Looking Telescope, first articulated in the NCRIS National Research Infrastructure Road Map [2], requires a coordinated data and analytics platform, these five, identified, fairly major disconnects needed to be addressed. Further, the pre-2016 developments really needed to be modernised. In particular, python/Jupyter notebooks did not exist in 2007 and the 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. The ambition was to create a highly flexible online environment in which researchers could find and access data and tools as online services to compose their own workflows to suit their specific research needs. Hence, the GeoDeVL program offered an opportunity to lay the foundations of a modern integrated platform to support the Downward Looking Telescope.

KEY ACHIEVEMENTS OF THE GEODEVL PROJECTS

Each GeoDeVL project had four work packages: MT, PS, International Geo Sample Number (IGSN), and the AVRE platform. Each worked towards the foundations for the Downward Looking Telescope as follows:

1. Magnetotellurics (MT) Work Package

The MT Work package was initially a 'data rescue' effort and 26 datasets from The University of Adelaide were made accessible online from the NCI GeoNetwork Catalogue. A basic data ingestion to publication pipeline has been developed for the three main MT data types (Level 0-1: Time Series; Level 2: Frequency Domain Processed Transfer Functions (Electronic Data Interchange (EDI) files); and Level 3: derived model outputs): each can be repurposed for any academic research group wanting to publish MT data. The GeoDeVL Project also focused on making raw time series datasets more accessible online. Trial reprocessing of this time series data using improved HPC applications at NCI have led to increased flexibility and options in processing methodologies, resulting in considerable improvements in MT response functions.

2. Passive Seismic (PS) Work Package

The GeoDeVL projects only partially funded the work done on making the RSES ANU PS Data FAIR and available online through the AusPass Portal (http://auspass.edu.au/), and compliant with web services standards of the International Federation of Digital Seismograph Networks (FDSN). This has been a high quality data rescue initiative and considerable effort has been made to capture the original metadata from the field and in developing detailed data QA/QC processes. In recognition that this work is in line with best practice internationally, the AusPass services recently made the international list of Data centres supporting FDSN web services: https://www.fdsn.org/webservices/datacenters/ and is the only Australian site on this list (https://www.auscope.org.au/posts/auspass-fdsn).

3. IGSN Minting Service Work Package

Collaboration between ARDC, CSIRO, and the Australian Geoscience research community developed a minting service for the allocation of IGSN unique identifiers for physical samples (http://auscope.org.au/igsn-info). It is now part of the ARDC citation and identifier services for data, samples and software (http://ardc.edu.au/resources/working-with-data/citation-identifiers). A bulk loader was also developed. Researchers can now connect data to source samples.

4. AVRE Work Package

The focus was to try to rationalise the plethora of geoscience data discovery services, portals, and software development into an integrated platform that can support a cradle-to-the-grave approach for data acquisition, capture, model generation/assimilation through to publication and curation. This required attention to making software FAIR and in a collaboration with ARDC, the development of a metadata profile compliant with international best practice (e.g., CodeMeta). Work is still ongoing as how to best implement those standards and to potentially trial Zenodo, a tool that can, amongst others, archive software releases directly from Github and issue each release with a DOI to help in discovery and enable proper accreditation for software developers. Further, individual portal components were modularized and turned into loosely coupled services and libraries, which can be more easily integrated into specific endpoints and tools.

REFERENCES

1. Wilkinson, M.D., Dumontier, M., Aalbersberg, IJ.J., Appleton, G., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Nature Scientific Data 3. https://doi.org/10.1038/sdata.2016.18 Accessed 29 August 2019.
2. 2016 National Research Infrastructure Road Map. https://docs.education.gov.au/system/files/doc/other/ed16-0269_national_research_infrastructure_roadmap_report_internals_acc.pdf Accessed 29 August, 2019.

ACKNOWLEDGEMENTS

The Geosciences DeVL project was funded by the Australian Research Data Commons, AuScope, NCI, CSIRO, Research School of Earth Sciences and The University of Adelaide.