eResearch AUSTRALASIA 2019

AuScope Dirt to Assimilation to Publishing Platform (ADAPPt): Increasing the Time for Research for Solid Earth Scientists

Lesley Wyborn[1], Tim Rawling[2], Ben Evans[1], Ryan Fraser[3], Nigel Rees[1], Jens Klump[3], Carsten Friedrich[4]

[1]National Computational Infrastructure, ANU, Canberra, Australia, Firstname.Lastname@anu.edu.au
[2]AuScope Ltd, Melbourne Australia, tim.rawling@unimelb.edu.au
[3]Mineral Resource, CSIRO, Perth, Australia, Firstname.Lastname@csiro.au
[4]Data 61, CSIRO, Canberra, Australia, Carsten.Friedrich@data61.csiro.au

SUMMARY

Since 2006 AuScope has been the national provider of research infrastructure to the earth and geospatial sciences communities in Australia and has invested heavily in a diverse suite of data acquisition infrastructures, as well as the development of numerical simulation and inversion codes. In line with the 2016 National Research Infrastructure Roadmap, AuScope is now developing the next phase of research infrastructure focused around the concept of the Downward Looking Telescope, which will be a distributed observational, characterisation and computational infrastructure providing the capability for Australian geoscientists to image and understand the composition of the Australian Plate with unprecedented fidelity. This ambition requires a modern eResearch platform that has machine actionable access to data in its rawer forms, with applications/software accessible as online services. The platform will need to be able to integrate across the disciplines of Geology, Geophysics and Geochemistry and beyond to domains such as Marine, Environment, Climate, Water and Urban, and ultimately to the Social Sciences. The growing number of relevant data and software will require alignment with the Findable, Accessible, Interoperable and Reusable (FAIR) principles [1] to enable online machine to machine interactions between data, software and compute.

INTRODUCTION

AuScope is the national provider of research infrastructure to the earth and geospatial sciences communities. Funded through the NCRIS scheme, since 2007 it has invested heavily in a diverse suite of infrastructures ranging from VLBI telescopes to geochronology/geochemistry laboratories and national geophysical data acquisitions. Portals were built to access solid Earth science data and AuScope also funded development of numerical simulation and inversion codes.

However, the reality is that it is estimated that only a low percentage of Australian solid Earth science data collected with public funding is actually FAIR [1] and discoverable and accessible in machine-actionable formats as will be required by the Downward Looking Telescope. It is commonly quoted that researchers of today can spend up to 80% of their time either trying to manage their own data, software or samples or to find other data, software and samples that can be relevant to their research. New demands for data, software and samples collected as part of publicly funded research grants and/or cited in publications [2] to comply with the FAIR principles (i.e., be stored in trusted repositories over the longer term, have persistent identifiers and landing pages and for both data and metadata to be machine actionable) is cutting even more into valuable research time. All this adds up to high levels of frustration for solid Earth scientists, as the time available for research is further reduced. Yet many of the key elements which scientists require to comply with the requirements of funders and publishers are already operational in modern research environments.

THE KEY ELEMENTS AND REALITIES OF MODERN RESEARCH ENVIRONMENTS

Much of today's research is undertaken in the digital world - most data are born digital and can be stored online in the cloud or in distributed repositories. Theoretically, once a researcher is back at their desk, software can be accessed from online software repositories, whilst processing can be done locally, on the cloud, or on supercomputers without the researcher having to manage these resources physically within their department.

But solid Earth science researchers have been slow to fully exploit these opportunities: the reality is that their research goals are constrained by the lack of interoperability at the “soft infrastructure” layers: the major issues are both between the different data and software used by the subsurface (geophysics, geochemistry, geology, etc.) as well as different approaches across the research, government and industry sectors. Much of Australian Earth science data, software and samples cannot be discovered and/or accessed online, and increasingly if ‘data’ is available, it is only as highly evolved data products, often as lower resolution images. Even when a researcher can obtain the rawer data, incompatible formats, vocabularies and ontologies mean that valuable research time is spent formatting and wrangling the data to a common standard and then rewriting parts of the software to enable connection to the relevant data.

DATA, SOFTWARE AND COMPUTE INFRASTRUCTURES FOR THE DOWNWARD LOOKING TELESCOPE.

To make research underpinning the Downward Looking Telescope as efficient and effective as possible, we need to ensure relevant data are born connected to community agreed meta(data) standards and that any primary data and derived data products are FAIR: software and tools also need to be FAIR. Both data and software need to be accessible online via standardised interfaces and capable of being processed in a variety of environments ranging from on premise servers, cloud and supercomputers. Above all, at each stage along the Data-path, we need to ensure that there is attribution to all who contributed and proper licensing. But AuScope cannot do this in isolation within the Solid Earth Sciences: it will need to integrate with similar efforts in other domains. The National 2016 National Research Infrastructure Roadmap notes that "*Australia has the opportunity to build a "Integrated Data-Intensive Infrastructure" and … create a more integrated, coherent and reliable system to deal with the various needs of data-intensive, cross-disciplinary and global collaborative research*" [3].

MAKING A START: THE AUSCOPE VIRTUAL RESEARCH ENVIRONMENT

The AuScope Virtual Research Environment (AVRE) was launched in 2017, through the ARDC GeoDeVL program, as part of a broader strategic goal to develop a combined geoscientific data discovery and assimilation/analysis platform to support the Downward Looking Telescope. Based around three discipline-based networks: Geophysics, Geochemistry, and Geology, it currently comprises a suite of data catalogues/portals, Jupyter notebooks, online minting services, etc. AVRE has started to make it easier to access tools and data, but you have to know they exist in order to access them.

INTRODUCING THE AUSCOPE DIRT TO ASSIMILATION TO PUBLICATION PLATFORM (ADAPPT)

We now need to link the components of AVRE into an integrated platform that will enable interdisciplinary research both within and beyond the solid Earth sciences and leverage common elements. Although each network (Geophysics, Geochemistry and Geology) serves different user communities and uses different international standards, all three follow a similar Dirt-to-Assimilation-to-Publication pattern and hence, the supporting eResearch infrastructures required for acquiring/deriving, discovering and accessing data and software is similar: all require access to storage and compute of varying capacities. Each network also requires the ability for their researchers to make their input artefacts (data, software, samples) compliant with new requirements from funders and publishers.

1) The Dirt Phase

All 3 networks either collect samples in the dirt or take measurements in the dirt: the primary data for all 3 requires a suite of software tools for calibration and processing. For the Dirt Phase, a generic tool and data logger infrastructure could be built for field data acquisition, using a different template depending on the data being collected and tailored to the workflow in the field. All data loggers in Australia need to be robust to operate effectively in the harsh conditions of the outback and intermittent internet connectivity: this development could be shared. Likewise, where samples are analysed in laboratories there are common patterns related to capturing data from instruments.

2) The Assimilation Phase

Data processing, assimilation, simulation and modelling all require the ability to discover and access data and software online. If the (meta)data complies with agreed community standards across the solid Earth sciences, then tools can be written to that standard without the need to either reformat the data, or change the software. For academics there is an increasing demand for researchers to have access to flexible processing environments that allow experimentation and support innovation, rather than rigid, fixed workflows. These require a standardised way of discovering and accessing software and tools and deploying them to compute, though the supporting platform.

3) The Publication Phase

The publication phase will be less tiresome for the researcher because at the Dirt Phase, their data has already been deposited in a repository in standardised formats with agreed metadata. Further through judicious use of identifiers and proper documentation of data and software as it is developed/used along the Data-path, the researcher already has the provenance of the input artefacts automatically recorded.

CONCLUSION

The potential is there for the solid Earth community to build an integrated eResearch platform that captures data along multiple Data-paths from Dirt to Assimilation to Publication and enable researchers to efficiently discover and access online data, software and compute resources, and above all, increase their effective research time. However, the solid Earth community does need to work carefully through this, to understand how to build its data networks to be compatible with those of other communities, whilst at the same time preserving the quality and integrity of the data within each of the key solid Earth domains: geophysics, geochemistry and geology.

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. Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., and Wyborn, L., (2019). Making Scientific Data FAIR. Nature **570**, 27-29 https://doi/10.1038/d41586-019-01720-7 Accessed 29 August 2019.
3. Australian Government (2016). National Research Infrastructure Roadmap. https://docs.education.gov.au/system/files/doc/other/ed16-0269_national_research_infrastructure_roadmap_report_internals_acc.pdf Accessed 29 August 2019.