EGU22-11012, updated on 14 Aug 2022
https://doi.org/10.5194/egusphere-egu22-11012
EGU General Assembly 2022

The Known Knowns, the Known Unknowns and the Unknown Unknowns of Geophysics Data Processing in 2030

Lesley Wyborn[1], Nigel Rees[2], Jens Klump[3], Ben Evans[4], Tim Rawling[5], and Kelsey Druken[6]
[1]Australian National University, National Computational Infrastructure, Acton, Australia (lesley.wyborn@anu.edu.au)
[2]Australian National University, National Computational Infrastructure, Acton, Australia (nigel.rees@anu.edu.au)
[3]Australian National University, National Computational Infrastructure, Acton, Australia (ben.evans@anu.edu.au)
[4]CSIRO Mineral Resources, CSIRO, Perth, Australia (jens.klump@csiro.au)
[5]Auscope Ltd, University of Melbourne, Melbourne, Australia (tim@auscope.org.au)
[6]Australian National University, National Computational Infrastructure, Acton, Australia (kelsey.druken@anu.edu.au)

The Australian 2030 Geophysics Collections Project seeks to make accessible online a selection of rawer, high-resolution versions of geophysics datasets that comply with the FAIR and CARE principles, and ensure they are suitable for programmatic access in HPC environments by future 2030 next-generation scalable, data-intensive computation (including AI and ML). The 2030 project is not about building systems for the infrastructures and stakeholder requirements of today, rather it is about positioning geophysical data collections to be capable of taking advantage of next generation technologies and computational infrastructures by 2030.

There are already many **known knowns** of 2030 computing**:** high end computational power will be at exascale and today's emerging collaborative platforms will continue to evolve as a mix of HPC and cloud. Data volumes will be measured in Zettabytes (10^{21} bytes), which is about 10 times more than today. It will be mandatory for data access to be fully machine-to-machine as envisaged by the FAIR principles in 2016. Whereas we currently discuss Big Data Vs (volume, variety, value, velocity, veracity, etc), by 2030 the focus will be on Big Data Cs (community, capacity, confidence, consistency, clarity, crumbs, etc).

So often today's research is undertaken on pre-canned, analysis-ready datasets (ARD) that are tuned towards the highest common denominator as determined by the data owner. However, increased computational power colocated with fast-access storage systems will mean that geophysicists will be able to work on less processed data levels and then transparently develop their own derivative products that are more tuned to the parameters of their particular use case. By 2030, as research teams analyse larger volumes of high-resolution data they will be able to see the quality of their algorithms quickly and there will be multiple versions of open software being used as researchers fine tune individual algorithms to suit their specific requirements. We will be capable of more precise solutions and in hazards space and other relevant areas, analytics will be done in faster-than-real-time.

The **known unknowns** emerging are how we will preserve and make transparent any result from

this diversity and flexibility with regards to the exact software used, the precise version of the data accessed, and the platforms utilised, etc. When we obtain a scientific 'product', how will we vouch for its fidelity and ensure it can be consistently replicated to establish trust? How do we preserve who funded what so that sponsors can see which investments have had the greatest impact and uptake?

To have any confidence in any data product, we will need to have transparency throughout the whole scientific process. We need to start working now on more automated systems that capture provenance through successive levels of processing, including how it was produced and which dataset/dataset extract was used. But how do we do this in a scaleable, machine readable way?

And then there will be the **unknown unknowns** of 2030 computing. Time will progressively expose these to us in the next decade as the scale and speed at which collaborative research is undertaken increases.