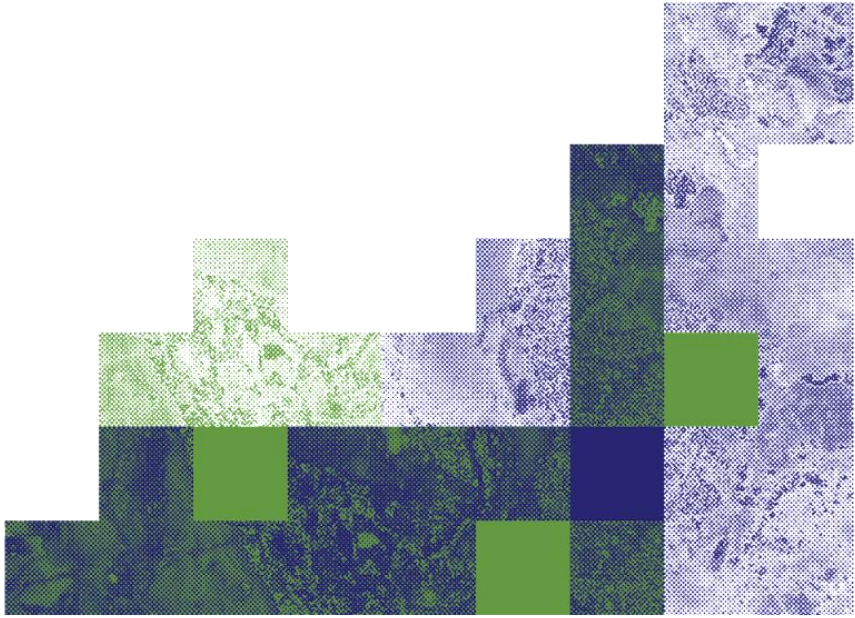
Open Science in Data-intensive Research Requires Multiple Entry Points: A Case Study from AuScope in Solid Earth Science Infrastructures.

:

Lesley Wyborn | AuScope, NCI, ARDC, ANU
Rebecca Farrington | AuScope
Alex Hunt | CSIRO Minerals
Jens Klump | CSIRO Minerals
Bryant Ware | Curtin University
Angus Nixon | University of Adelaide



John de Laeter Centre Curtin University | THE UNIVERSITY of ADELAIDE | AuScope Geochemistry Network | NCI AUSTRALIA | CSIRO | Australian National University | AuScope | NCRIS

We acknowledge the Traditional Owners of the land on which our research infrastructure and community operate across the Australian continent, and pay our respects to Elders past and present.

We recognise the connection they have with land, sea, sky and waterways for tens of thousands of years.

ABSTRACT

Solid Earth Science datasets provide evidence-based insights into surface and subsurface environments, including the quantification of longitudinal changes over decades. However, their increasing diversity and scale present significant challenges. Primary Observational Datasets (PODs) range widely in size, from small-scale collections in the megabyte range, suitable for on-premise or cloud storage, to high-volume collections that are petabytes in volume and require co-located High Performance Compute-Data (HPC-D) platforms for timely, effective analysis.

Many funders now request compliance with the FAIR, CARE and TRUST principles, whilst increasing demands for Open Science set a high bar for reproducibility, transparency and sharing requiring open publication of all data collected, tools and processes (UNESCO, 2021; https://doi.org/10.5281/zenodo.5741832).

It is no longer possible for a single repository to meet these requirements and serve all users, who range from expert power-users to novices. Instead, a 'Repository Ecosystem' is needed, one that balances resources along the full-path of research data use, including:

1. Curation and sustainable preservation of raw full-resolution PODs captured directly off instruments;
2. Calibration and conversion of raw PODs into full-resolution reference datasets using community-agreed machine-readable data formats, standards and vocabularies and annotation with rich machine-actionable metadata;
3. Systematic reprocessing of PODs into reusable downstream analysis-ready products that meet specific researcher needs.

This paper outlines Auscope's approach to developing a Solid Earth Science Data Ecosystem that enables seamless access to PODs, hosted on HPC-D platforms and cloud environments, and clear pathways that connect these datasets to processed, analysis-ready data products delivered through distributed data platforms and portals.

What do we mean by Open Science?

—Open science is a set of principles and practices that aim to make scientific research from all fields accessible to everyone for the benefits of scientists and society as a whole.

—Open science is about making sure that not only scientific knowledge is accessible but also that the production of that knowledge itself is inclusive, equitable and sustainable.

—In other words, **Open Science** it is not just about the final data products needing to be open, we need to expose all steps along the pathway from the initial data collection point through generations of multiple derivative datasets.

UNESCO and Canadian National Commission for UNESCO, 2022. An Introduction to the UNESCO Recommendation on Open Science. https://doi.org/10.54677/XOIR1696

Introducing PODs…

- **Primary Observation Datasets**, or **PODs**, are the fundamental observations and analyses underpinning the published data

- PODs are **not traditionally stored or reported** as part of publications or data release, yet allow for the full recreation and reinterpretation of end-use data sets
  - advancing software, updated constants or reduction procedures
  - leveling of long-term compilations or multiple sources

- PODs are the ‘seeds’ of our future research

Image Sourrce
https://media.istockphoto.com/id/1005962176/photo/fresh-green-peas-isolated-on-white-background.jpg?s=612x612&w=0&k=20&c=FQd1HP9P70xud-1l8gbhVRSWUdGNfak-sLLH2WZ_0as=

Text based on slide of Angus Nixon

Why is it important to preserve PODs

The 2019 Beijing Declaration on Research Data asserts that **publicly funded research data** :

- **‘are, by default, in the public interest and should be accessible to the greatest extent possible for international reuse’.**
- **‘should be interoperable, and preferably without further manipulation or conversion, to facilitate their broad reuse in scientific research**’.

COMMITTEE ON DATA CODATA INTERNATIONAL SCIENCE COUNCIL

The Beijing Declaration on Research Data

Preamble

Grand challenges related to the environment, human health, and sustainability confront science and society. Understanding and mitigating these challenges in a rapidly changing environment require data[1] to be FAIR (Findable, Accessible, Interoperable, and Reusable) and as open as possible on a global basis. Scientific discovery must not be impeded unnecessarily by fragmented and closed systems, and the stewardship of research data should avoid defaulting to the traditional, proprietary approach of scholarly publishing. Therefore, the adoption of new policies and principles, coordinated and implemented globally, is necessary for research data and the associated infrastructures, tools, services, and practices. The time to act on the basis of solid policies for research data is now.

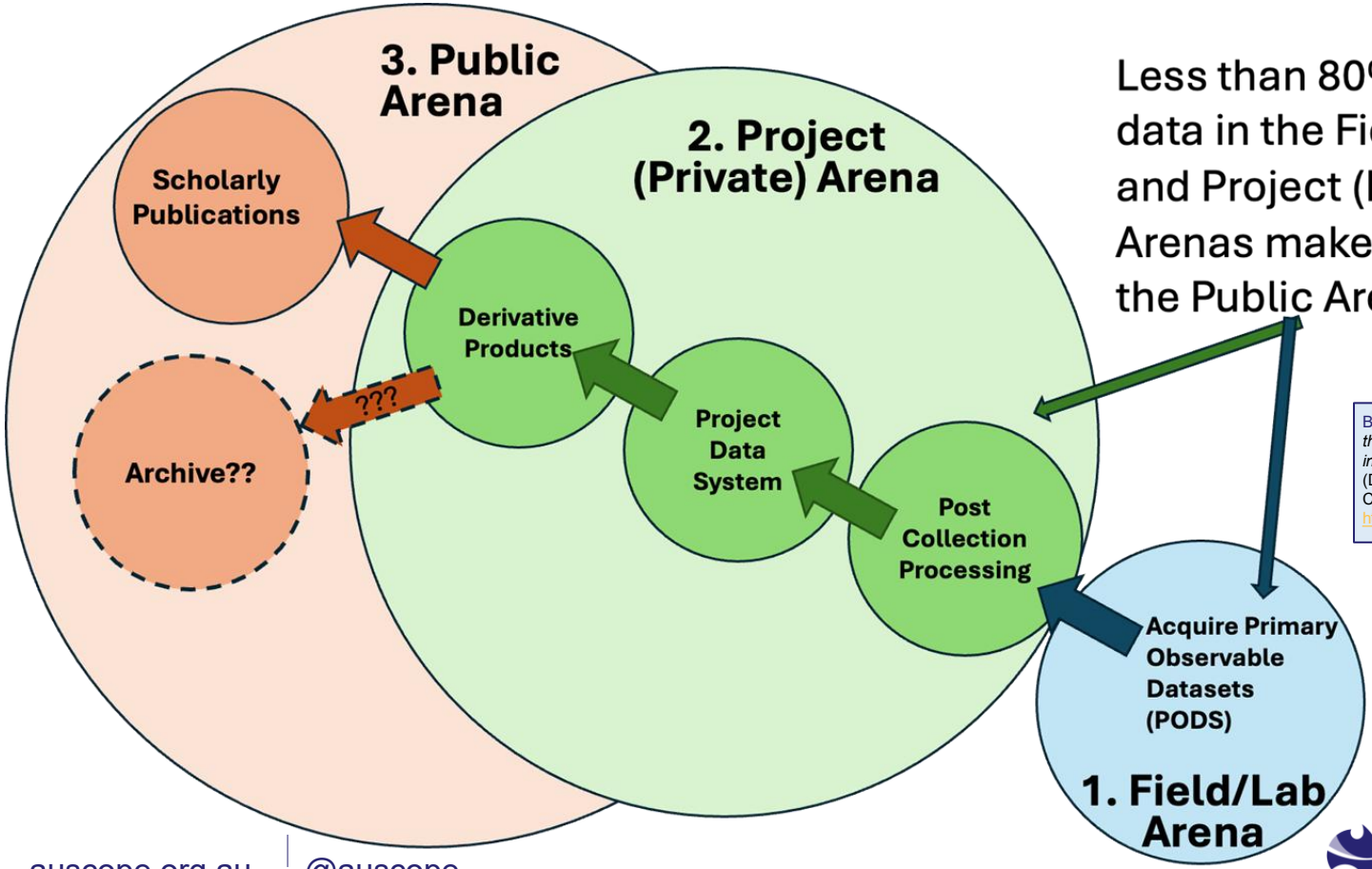
The Beijing Declaration is intended as a timely statement of core principles to encourage global cooperation, especially for public research data. It builds on and acknowledges the many national and international efforts that have been undertaken in the policy and technical spheres on a worldwide basis.[a] These major contributions are listed in the Appendix.

Several emergent global trends justify and precipitate this declaration of principles:

- Massive global challenges require multilateral and cross-disciplinary cooperation and the broad reuse of data to improve coherence concerning recent UN landmark agreements, such as the Paris Climate Agreement, the Sendai Framework for Disaster Risk Reduction, the Sustainable Development Goals (SDGs), the Convention on Biological Diversity, the Plant Treaty, the World Humanitarian Summit, and others. The comprehensive agendas for action provided by these agreements requires access to and reuse of all kinds of data.
- Research and problem-solving, especially addressing the SDG challenges, are increasingly complex and driven by ‘big data’, resulting in the need to combine and reuse very diverse data resources across multiple fields. This poses an enormous challenge in the interoperability of data and responsible stewardship, with full respect for privacy.
- Rapid advances in the technologies that generate and analyze data pose major challenges concerning data volume, harmonization, management, sharing, and reuse. At the same time, emerging technologies (including machine learning) offer new opportunities that require access to reusable data available in distributed, yet interoperable, international data resources.
- Changing norms and ethics encourage high-quality research through greater transparency, promote the reuse of data, and improve trustworthiness through the production of verifiable and reproducible research results. Increasing the openness of research data is efficient, improving the public return on investment, and generating positive externalities.
- Open Science initiatives are emerging globally, including in less economically developed countries. There consequently are opportunities for these countries to take advantage of technological developments to develop a greater share in scientific production. Without determined action, there is also a risk that the divide in scientific production will widen.

CODATA, C. on D. of the I. S. C., CODATA International Data Policy Committee, CODATA and CODATA China High-level International Meeting on Open Research Data Policy and Practice, Hodson, S., Mons, B., Uhlir, P., & Zhang, L. (2019). The Beijing Declaration on Research Data. Zenodo. https://doi.org/10.5281/zenodo.3552330

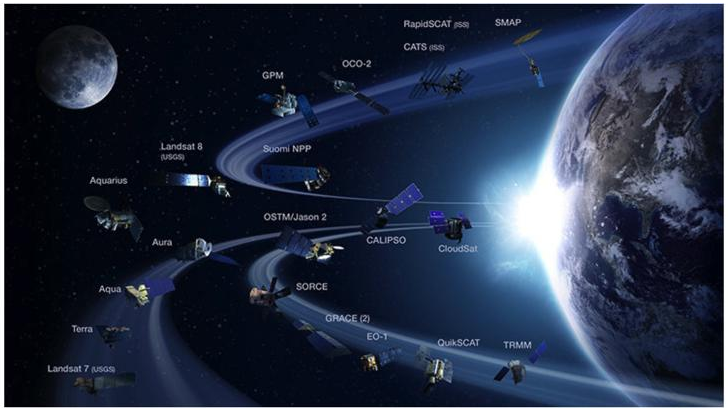
Traditional Research Methodologies are Linear and not Always Open



Less than 80% of data in the Field/Lab and Project (Private) Arenas make it to the Public Arena

Based on Baker, K.S., 2017. *Data work configurations in the field-based natural sciences: mesoscale infrastructures, project collectives, and data gateways* (Doctoral dissertation, University of Illinois at Urbana-Champaign) (Figure 6.1, p. 87) https://www.ideals.illinois.edu/items/103276

The NASA Processing Levels



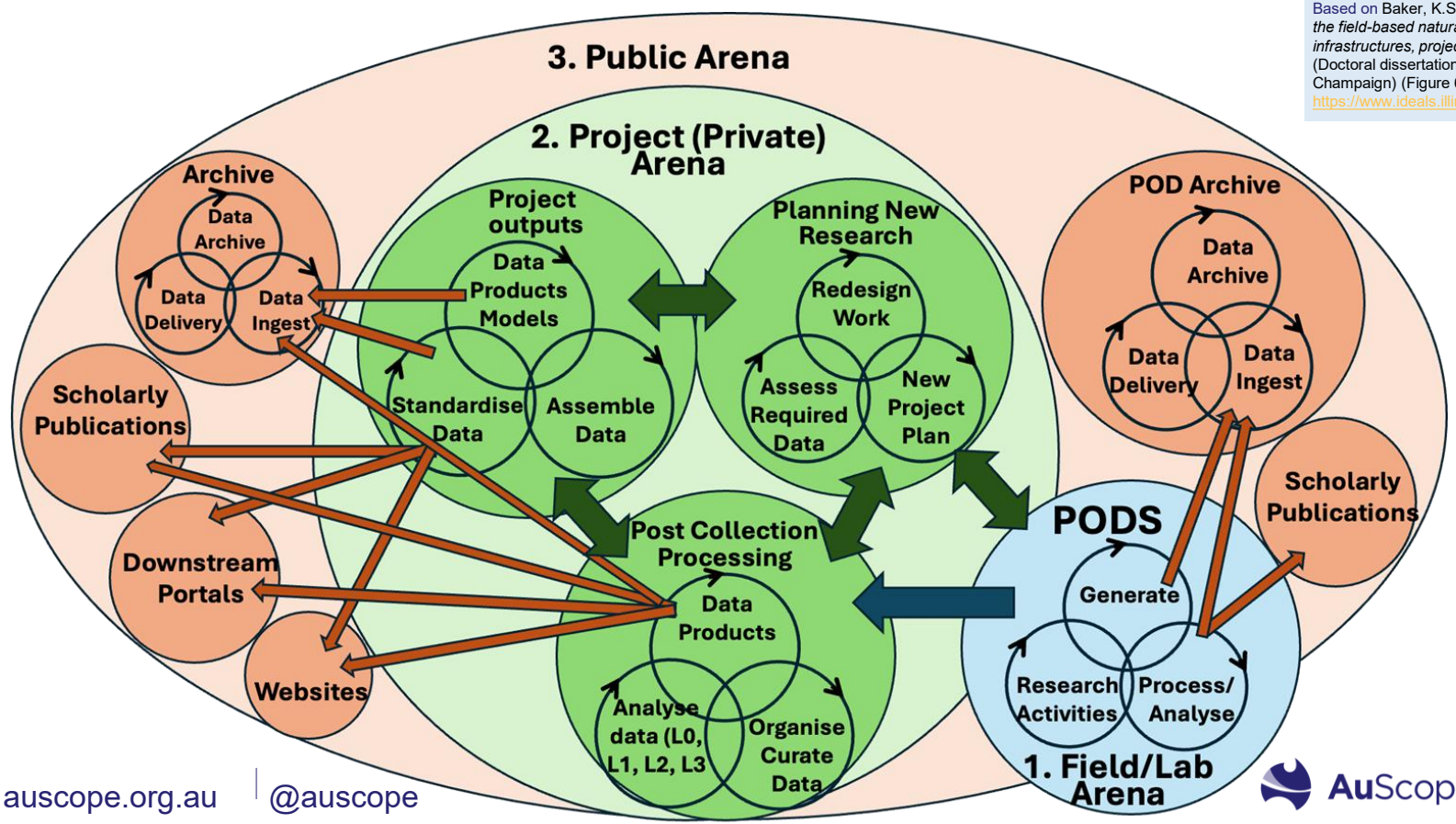
Source of Graphic: https://earthobservatory.nasa.gov/blogs/earthmatters/2015/04/29/elusive-earthquake-imagery/

- L0 = Reconstructed, unprocessed instrument data at full resolution
- L1 = L0 data time-referenced, annotated & processed to sensor units
- L2 = Derived geophysical variables at the same resolution
- L3 = Variables mapped onto uniform space-time grid scales
- L4 = Model outputs or results from analyses of lower-level data

Source: https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels

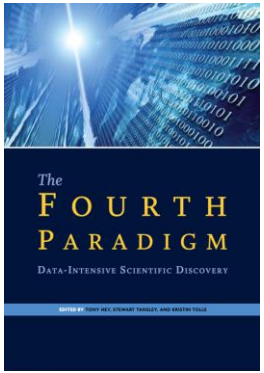
The Modern Open Data Research Data Cycle is Complex, but Enables Curation and Preservation of PODS

Based on Baker, K.S., 2017. *Data work configurations in the field-based natural sciences: mesoscale infrastructures, project collectives, and data gateways* (Doctoral dissertation, University of Illinois at Urbana-Champaign) (Figure 6.2, p 88) https://www.ideals.illinois.edu/items/103276



What do we mean by Data-Intensive Science?

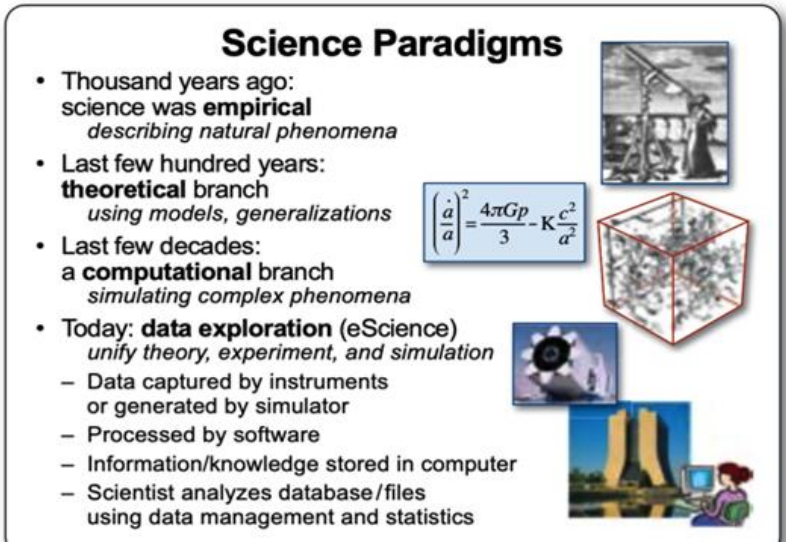
HHey, T., Tansley, S. and Tolle, K.M., 2009. Jim Gray on eScience: a transformed scientific method. *URL http://research. microsoft. com/en-us/collaboration/fourthparadigm/4th paradigm book jim gray transcript. pdf.*



— The term emerged in The Fourth Paradigm ~2009, inspired by Jim Gray, in which scientific progress is expected to stem from the discovery of patterns in huge volumes data.

— Data-intensive Science describes research and engineering efforts which depend on the processing of large datasets which are acquired from instruments such as cameras, genome sequencers, super microscopes and other devices.

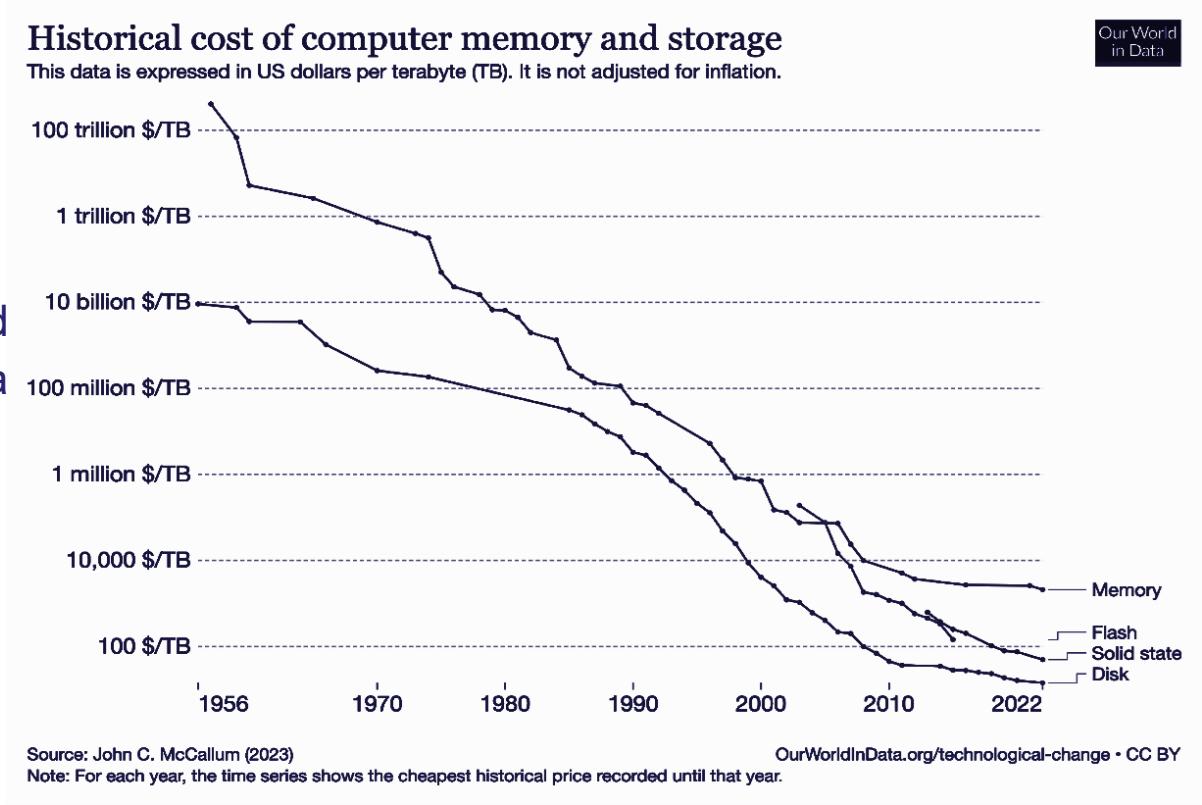
— AI and ML soon emerged as enablers of Data-Intensive Science

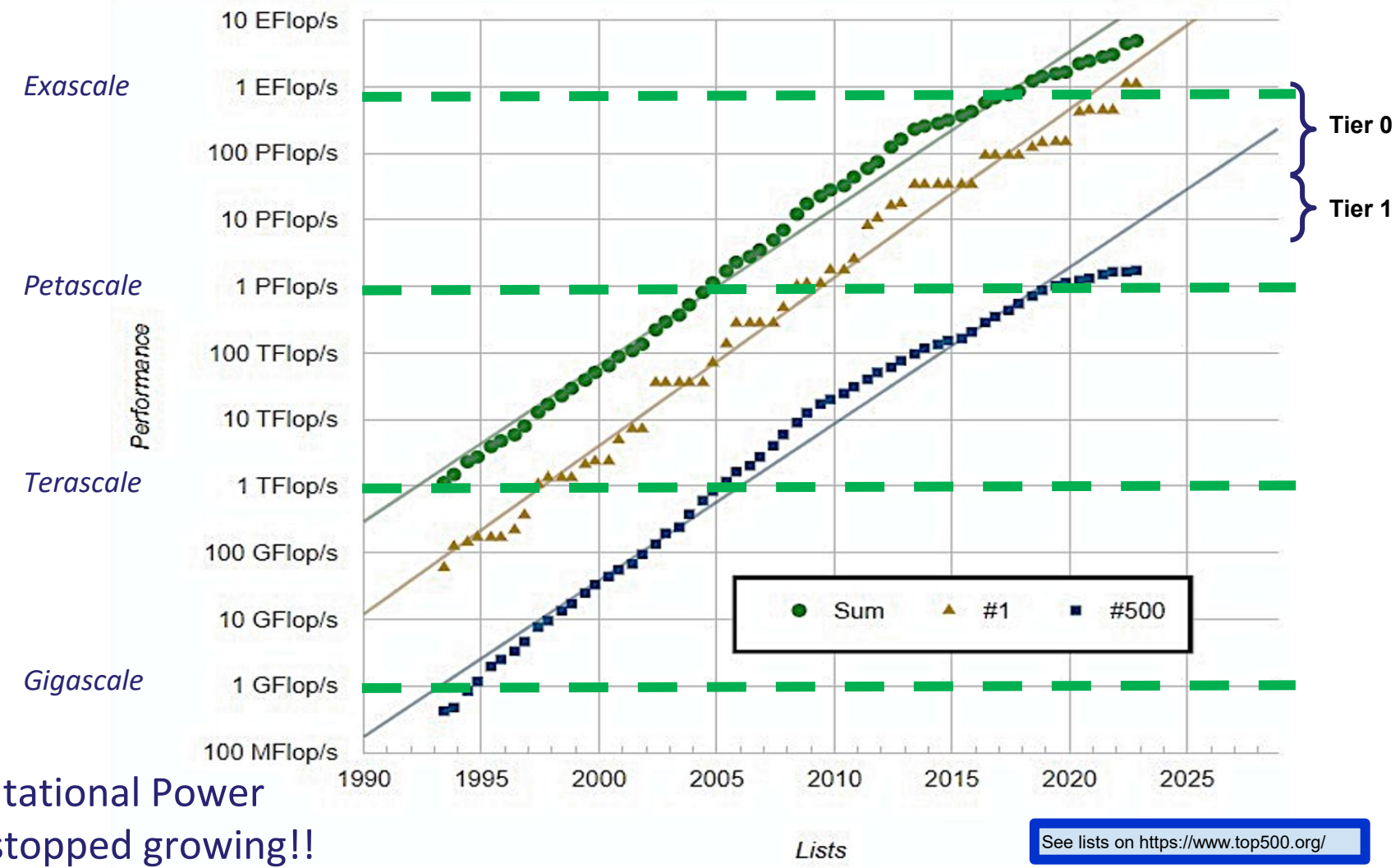


Why did it not happen sooner? in the 1990's storage and memory were expensive

- In 1980 storage was ~$1M per gigabyte
- For file based data, we could just could not have most data online at full resolution.

https://ourworldindata.org/technological-change

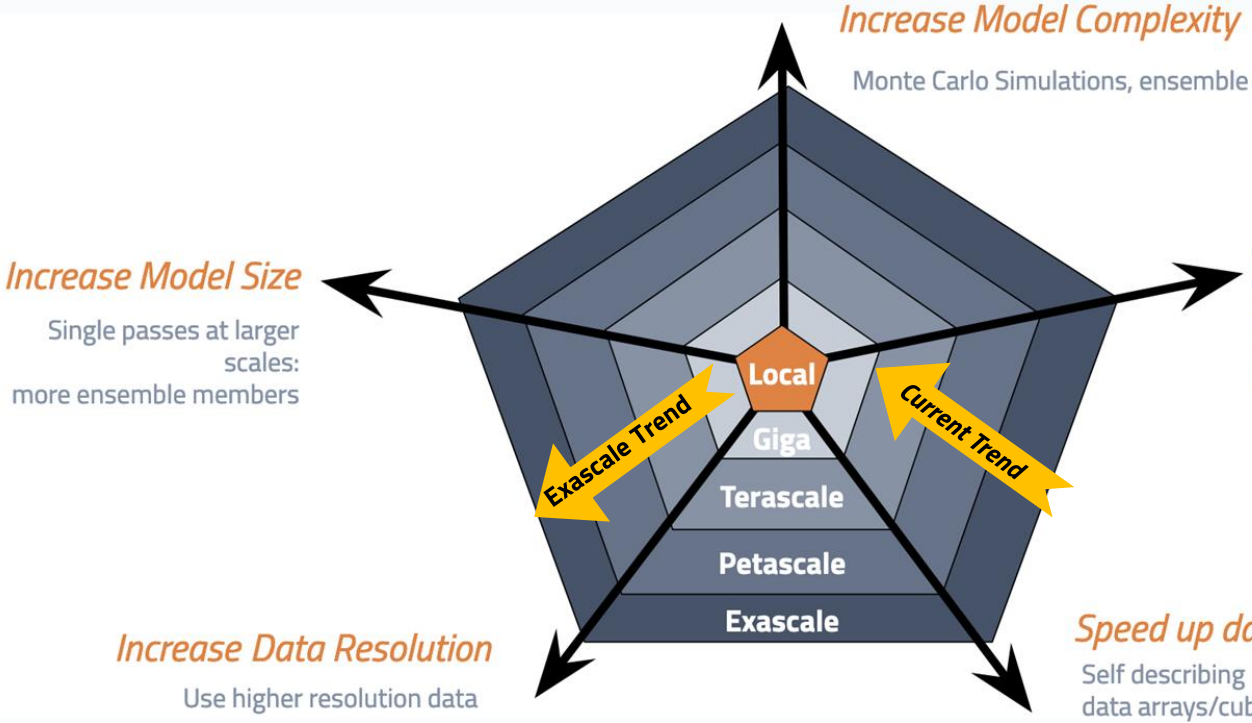




Computational Power never stopped growing!!

See lists on https://www.top500.org/

Scaling to Exascale: we need to separate storage and processing from distrubution

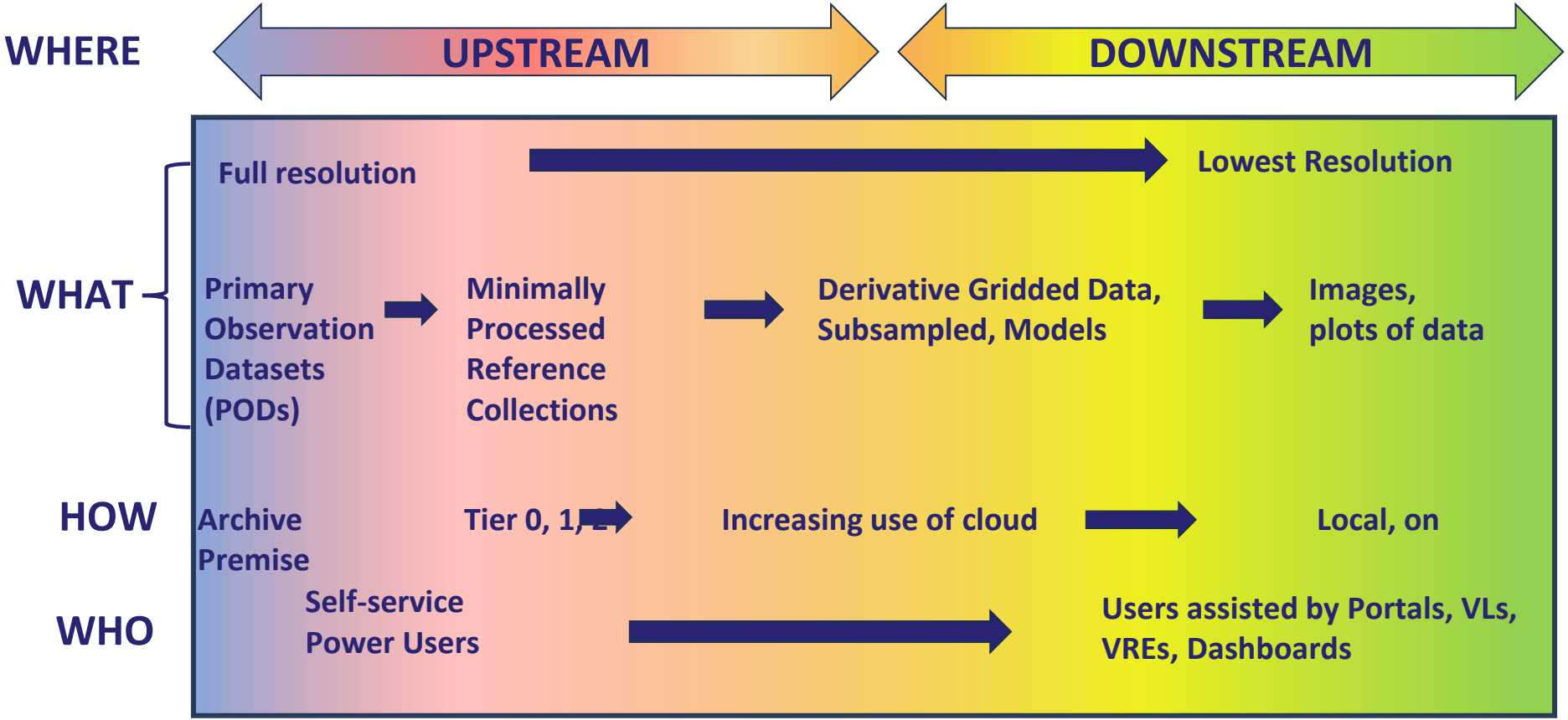


- We need to enable access to data at all levels (L0-L3)
- Science is often Bonsai-ed!
- One RI/infrastructure can not fit all demands

http://www.phoenixbonsai.com/SmallLogo.jpg

Adapted from: Wyborn, L., 2023: https://docs.google.com/presentation/d/17OIunofp5EpxNJ0iq8Z8YLqK2KvF-uZx/edit#slide=id.p1

Open Science enables catering for both Upstream & Downstream Users



AI is driving need for access to Upstream data

WHERE

UPSTREAM ← → **DOWNSTREAM**

WHAT

Full resolution → Lowest Resolution

Primary Observation Datasets (PODs) → Minimally Processed Reference Collections → Derivative Gridded Data, Subsampled, Models → Images, plots of data

AI is driving the demand for access to upstream standardized machine-actionable data

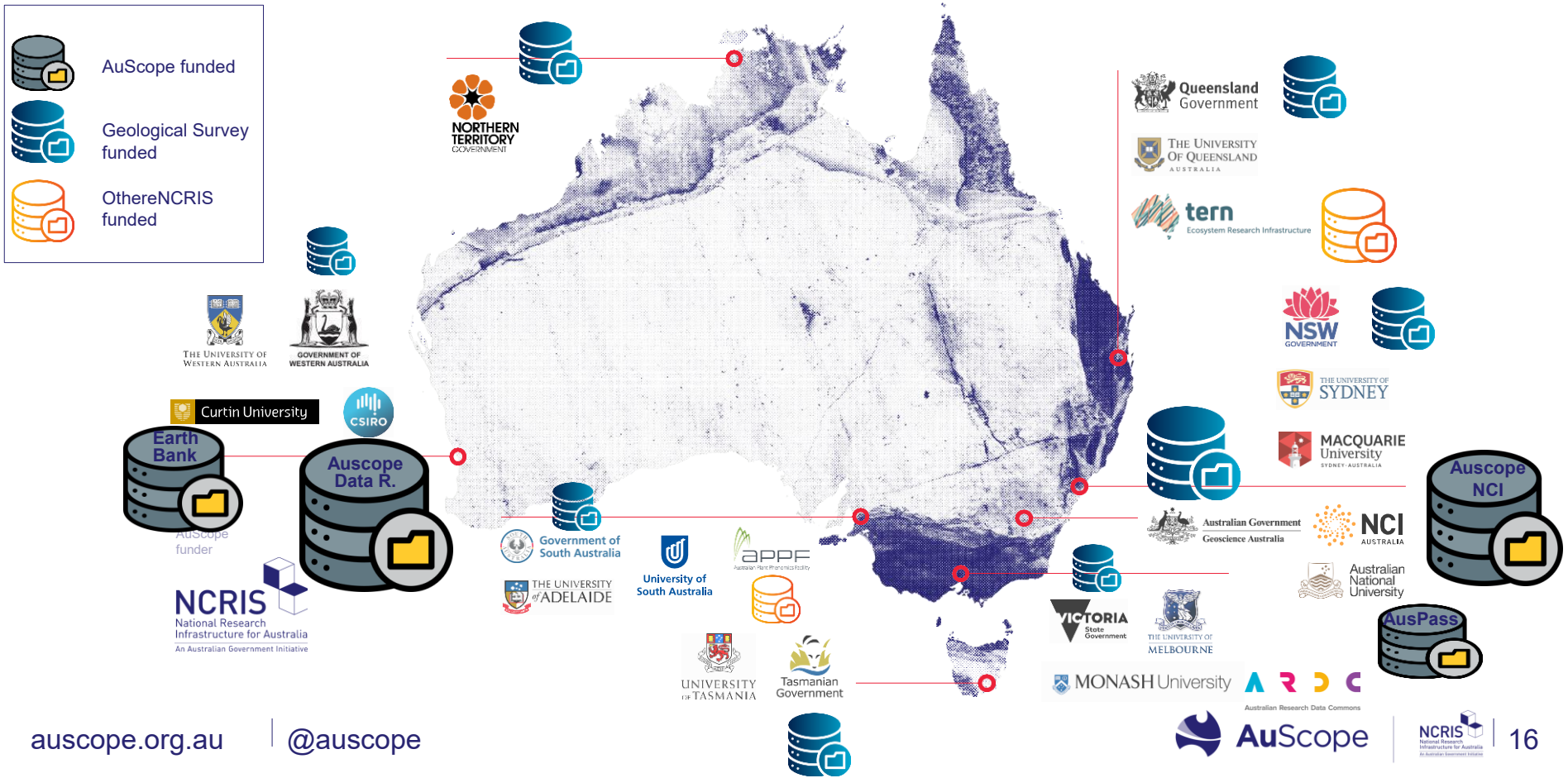
HOW

Archive Premise → Tier 0, 1, 2 → Increasing use of cloud → Local, on

WHO

Self-service Power Users → Users assisted by Portals, VLs, VREs, Dashboards

Australian Earth & Environmental Science **Curated** Repositories



AuScope Repository Landscape

AuScope Data Repository based at CSIRO Minerals Perth

- AWS based
- Designed more for Long Tail Collections
- Offers Independent backups for some of the institution data stores
- Has file size limit of 10 GB

AuScope Data Repository at NCI

- Co-located with HPC/Cloud
- Designed for large volume collections GBs, TBs, PB (geophysics, NVCL, Drone, Models)
- Can be sub-setted and downloaded locally

EarthBank based at Curtin, on AWS

- Specialist geochemistry database

AusPass based at ANU RSES

- Specialist Passive Seismic Collection

The basics of MT time series acquisition and processing

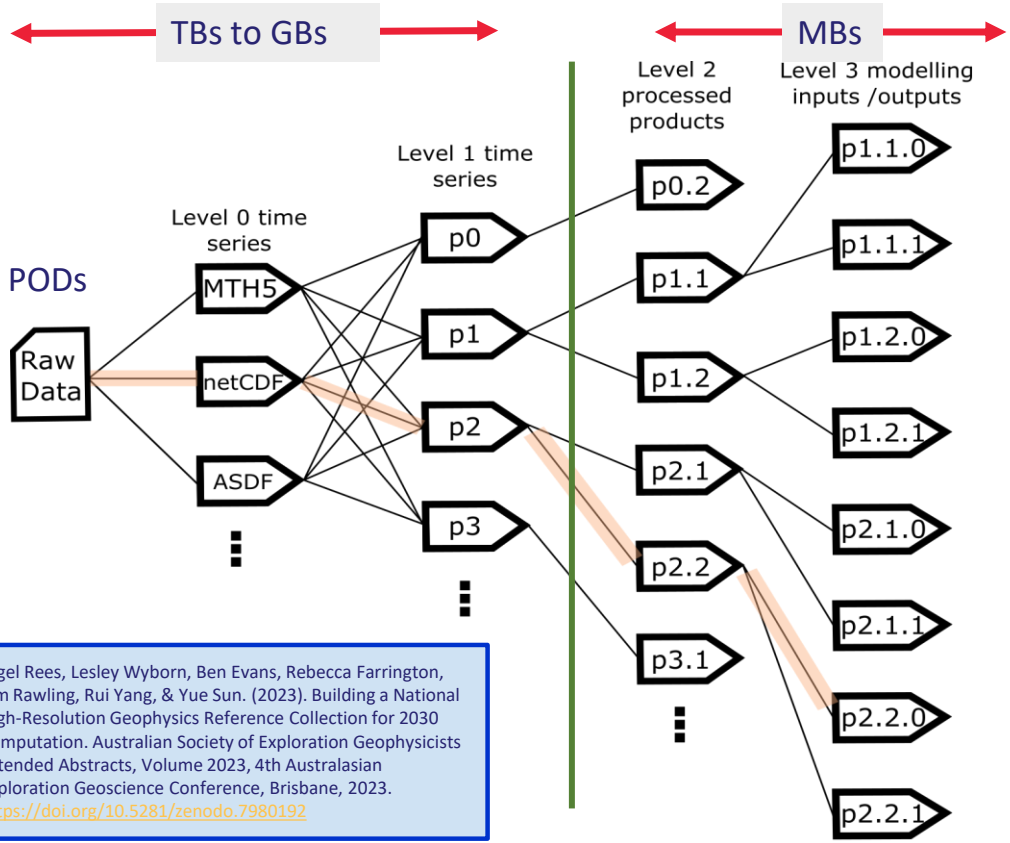
The **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**MTtsdp**) codes: https://github.com/nci/MTtsdp

Processing Levels	Name	Typical Volumes	Description
Packed Raw Data	Raw Time Series	GBs to TBs	Telemetry data streamed from site loggers
Level 0	Edited Time Series	GBs to TBs	Time ordered instrument recorded data (e.g., raw voltages, counts) at full resolution
Level 1	Transformed Time Series	GBs to TBs	Level 0 data that have been transformed (e.g., calibrated, resampled, rotated, had noisy data removed, filters applied).
Level 2	Derived frequency domain processed data	MBs	Geophysical parameters (e.g., impedance tensors) derived from frequency domain time series processing of Level 1 data
Level 3	Derived modelling inputs and outputs	MBs	Level 2 parameters converted into input files for modelling and inversion algorithms with outputs mapped onto space-time grids.

Rees, N., Evans, B., Heinson, G., Conway, D., Yang, R., Thiel, S., Robertson, K., Druken, K., Goleby, B., Wang, J., Wyborn, L. & Seillé, H., 2019. The Geosciences DeVL Experiment: new information generated from old magnetotelluric data of The University of Adelaide on the NCI High Performance Computing Platform, ASEG Extended Abstracts, 2019:1, 1-6, DOI: 10.1080/22020586.2019.12073015

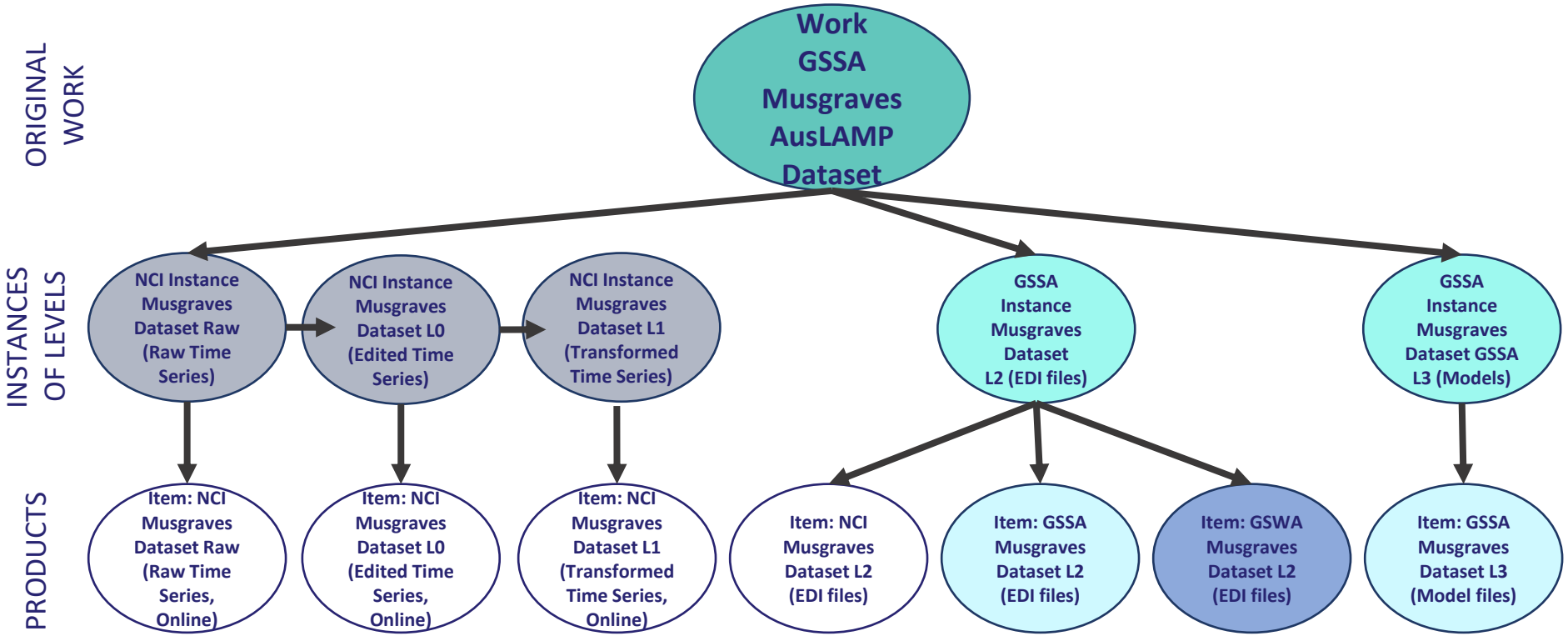
Vertical integration between processing levels through PIDs

Dataset of 93 MT stations	Serial I/O	MPI based Parallel I/O (96 cores)
Level 0: one MTH5/mt_metadata file for all stations	~ 14 hours	~ 35 minutes
Level 0: one MTH5/mt_metadata file per station	~ 5 hours 47 minutes	~ 4 minutes
Level 1: one MTH5/mt_meta-data file/station	~ 49 minutes	~ 1.2 minutes
Level 2: one EDI file/station	~ 2 hours 30 minutes	~ 2 minutes

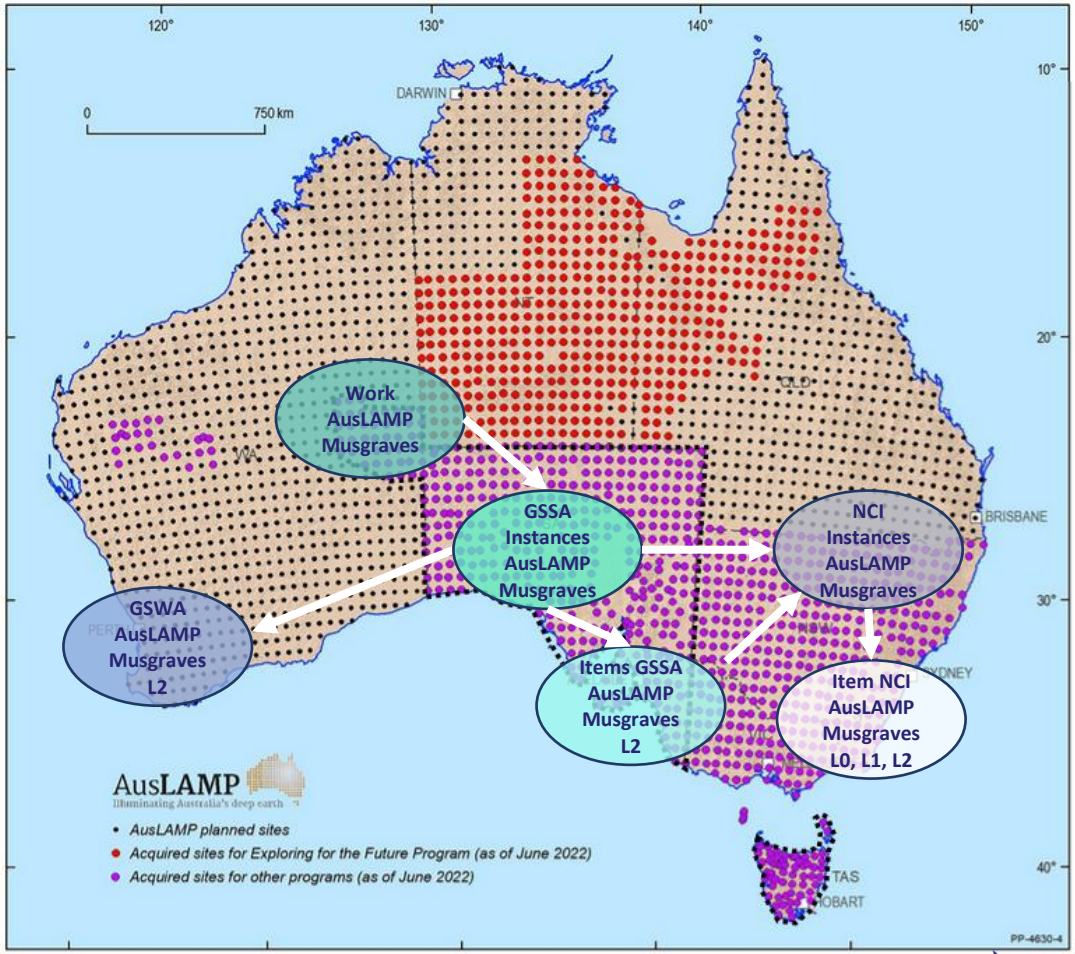


Nigel Rees, Lesley Wyborn, Ben Evans, Rebecca Farrington, Tim Rawling, Rui Yang, & Yue Sun. (2023). Building a National High-Resolution Geophysics Reference Collection for 2030 Computation. Australian Society of Exploration Geophysicists Extended Abstracts, Volume 2023, 4th Australasian Exploration Geoscience Conference, Brisbane, 2023. https://doi.org/10.5281/zenodo.7980192

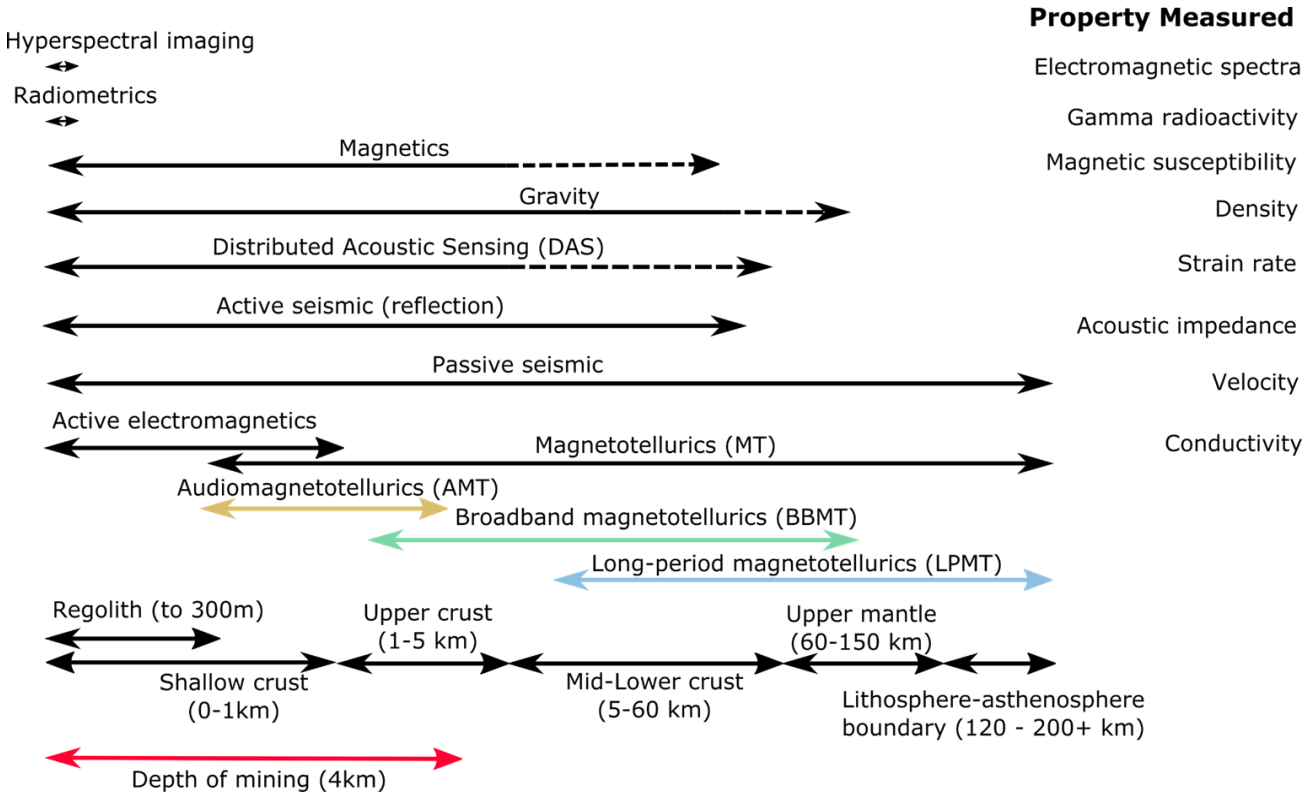
Vertical Integration of Levels of the AusLAMP Musgraves Dataset on multiple repositories



The AusLAMP Musgraves Dataset: who has got what where…



The Ambition: Multiphysics & Multi-scale HPC-D Geophysical Analysis

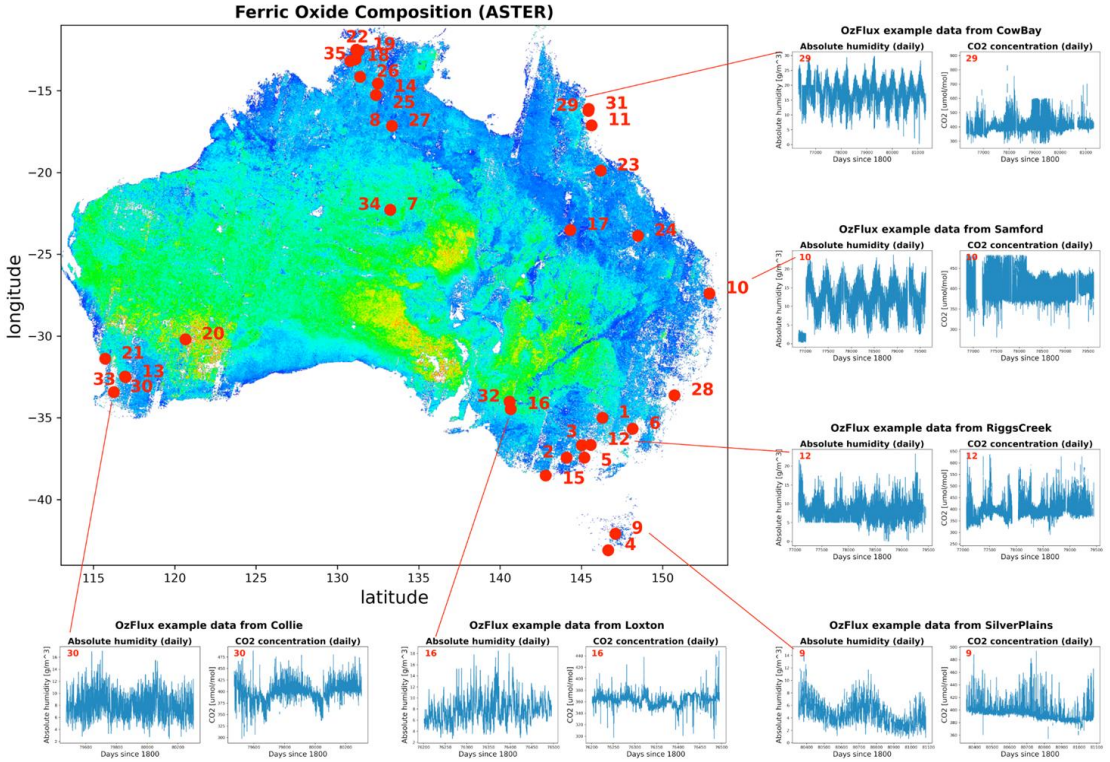


Types of geophysical data collected in Australia, the physical property measured and the depth of the crust that is sampled: also shown is the depth of current mining. Figure modified from original of Richard Chopping (GSWA).

Nigel Rees, Lesley Wyborn, Ben Evans, Rebecca Farrington, Tim Rawling, Rui Yang, & Yue Sun. (2023). Building a National High-Resolution Geophysics Reference Collection for 2030 Computation. Australian Society of Exploration Geophysicists Extended Abstracts, Volume 2023, 4th Australasian Exploration Geoscience Conference, Brisbane, 2023. https://doi.org/10.5281/zenodo.7980192

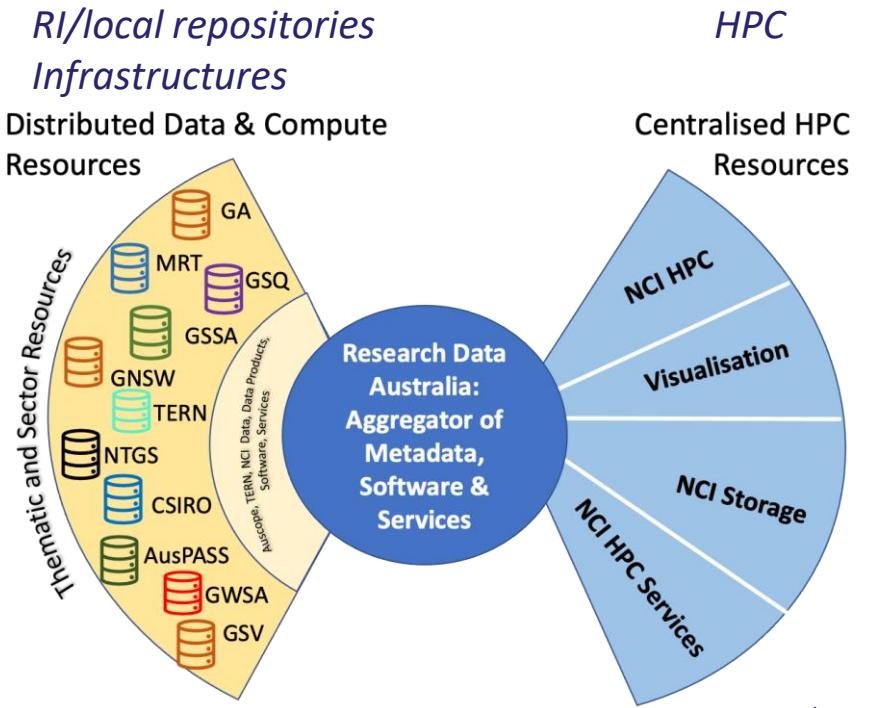
Horizontal Integration of Data from External Repositories on the fly

- Integrating a precise observational dataset to help calibrate at large volume ‘proxy’ dataset on HPC
- For example, external data (TERN’s OzFlux tower locations (red dots)) is integrated with data hosted at NCI (ASTER layer)



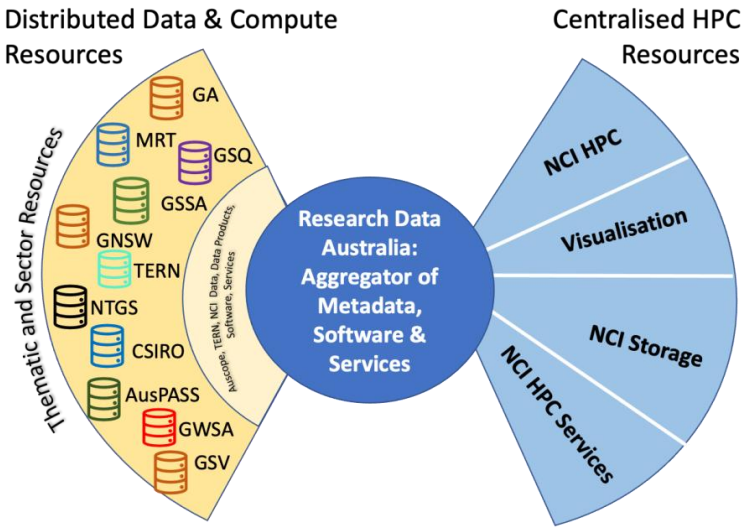
AuScope is enabling multiple entry points for varying skill levels

- Specialised data-intensive high resolution multiphysics processing can be done efficiently at NCI by expert users who have high levels of computer skills and just want to know how the data and the supporting software can be accessed.
- Lower volume and/or more highly processed downstream data products can be downloaded locally from other repositories and/or dashboards/ portals that facilitate more generally processing.
- Because Research Data Australia (RDA) harvests metadata (and PIDs) from many NCRIS RIs and Government Agencies, it is the middleware that enables researchers directly find and access the dataset at the level of processing of their interest.



Conclusions

- It is no longer possible for a single RI portal/dashboard/ platform to meet the needs and requirements of all researchers: we need multiple strategies and multiple points of entry.
- We need to ensure archiving, curation and access of the Primary observable datasets and enable in situ processing on large scale HPC/cloud environments by experts.
- For access to more commonly used lower resolution data products, individual RIs can focus on enabling access for local processing, downloads, etc, including websites/ platforms/ dashboards/notebooks that enable smaller scale processing online.
- We predict as data volumes increase, there will be a move towards more processing in situ: downloading data for local processing will decrease, particularly for multi-physics, data-intensive national/global scale research.
- Bonsai-ed Science will diminish!



Data is co-located with HPC/Cloud compute for in situ access

http://www.phoenixbonsai.com/SmallLogo.jpg