Preserving Primary Observable Datasets for Future Science

Session Organisers/Presenters:

Angus Nixon | University of Adelaide

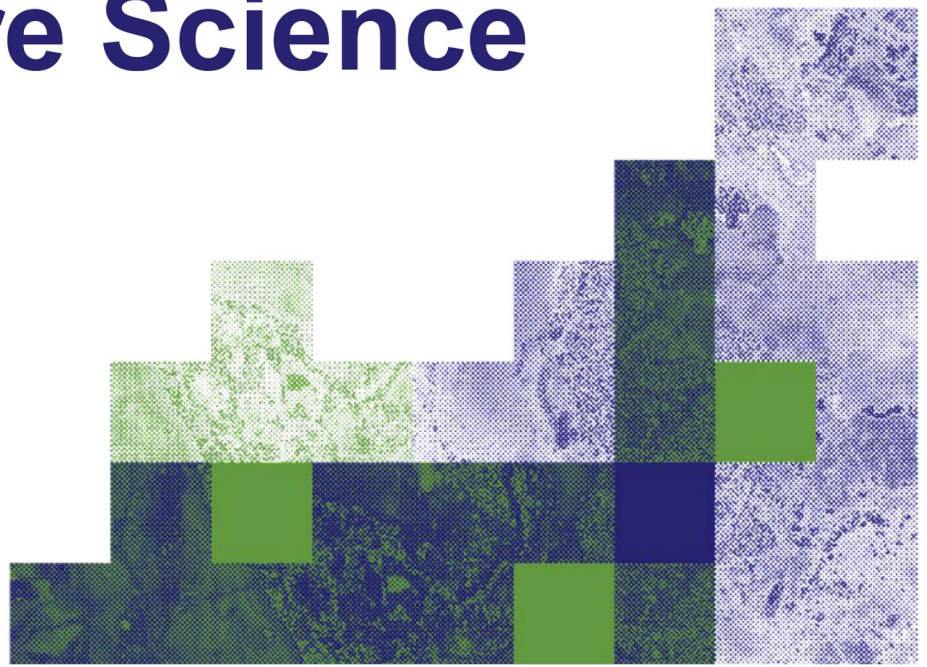
Bryant Ware | Curtin University

Lesley Wyborn | Australian National University

Session Presenters:

Simon Hodson | CODATA

Kelsey Druken | Australian National University



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.

BoF Outline

- **Introduction to PODs** - Angus Nixon (5 mins)
- **Cross Domain Interoperability Framework** - Simon Hodson (5 mins)
- **Domain Case Studies**
  - **Geochemistry** - Angus Nixon (5 mins)
  - **Geophysics** - Lesley Wyborn (5 mins)
  - **Climate** - Kelsey Druken (5 Mins)
- **Discussion** - all (30 mins)
  - **What kind of PODs does your community produce?**
  - **Are there initiatives to preserve PODs in your community?**
  - **What are the challenges for preserving PODs?**
- **Closing Remarks** - Angus Nixon (5 mins)

What Are PODs?

- **Primary Observable 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

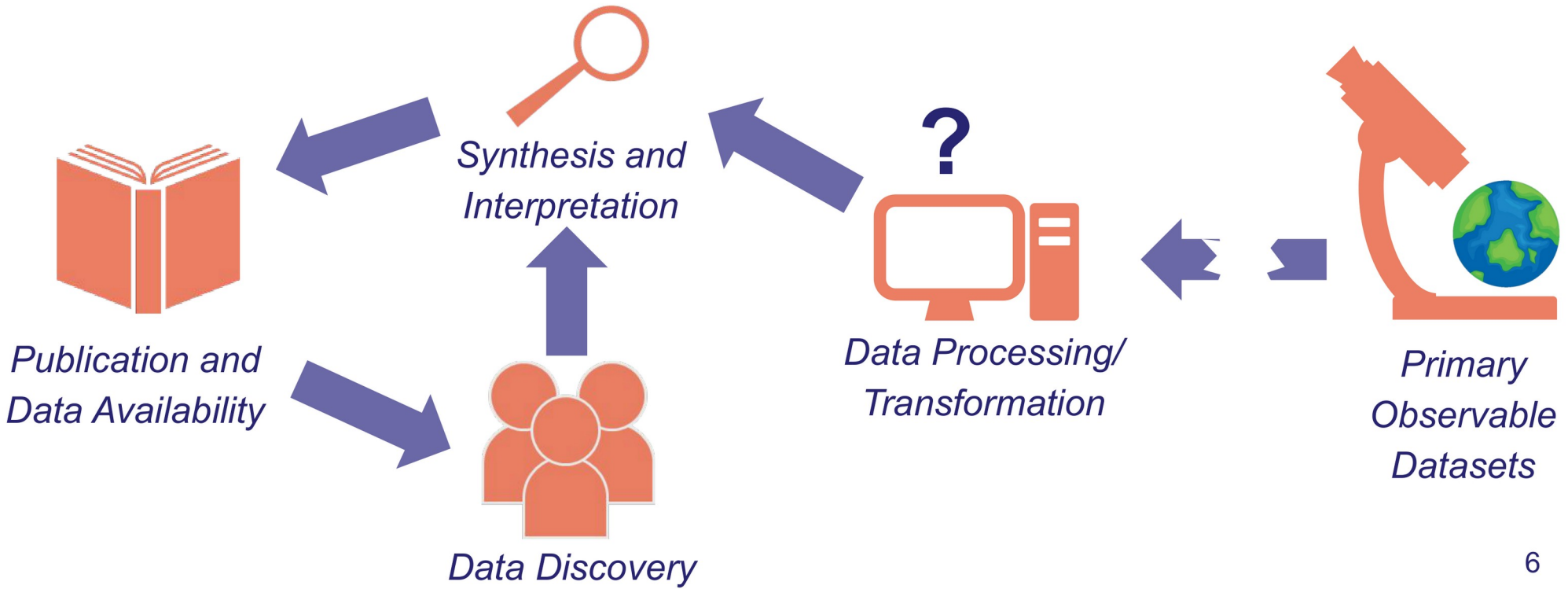
Scientific Data Processing Levels

Level 0: Reconstructed, unprocessed instrument or platform data at full resolution
Level 1A: Reconstructed, unprocessed instrument data at full resolution, with appended medatata
Level 1B: Level 1A data that have been processed to instrument units
Level 2: Derived variables at the same resolution and location as the Level 1 source data
Level 3: Data products derived from Level 2 and below data
Level 4: Model output or results from analyses of lower level data, e.g., variables derived from multiple measurements

After EOSDIS (2020)

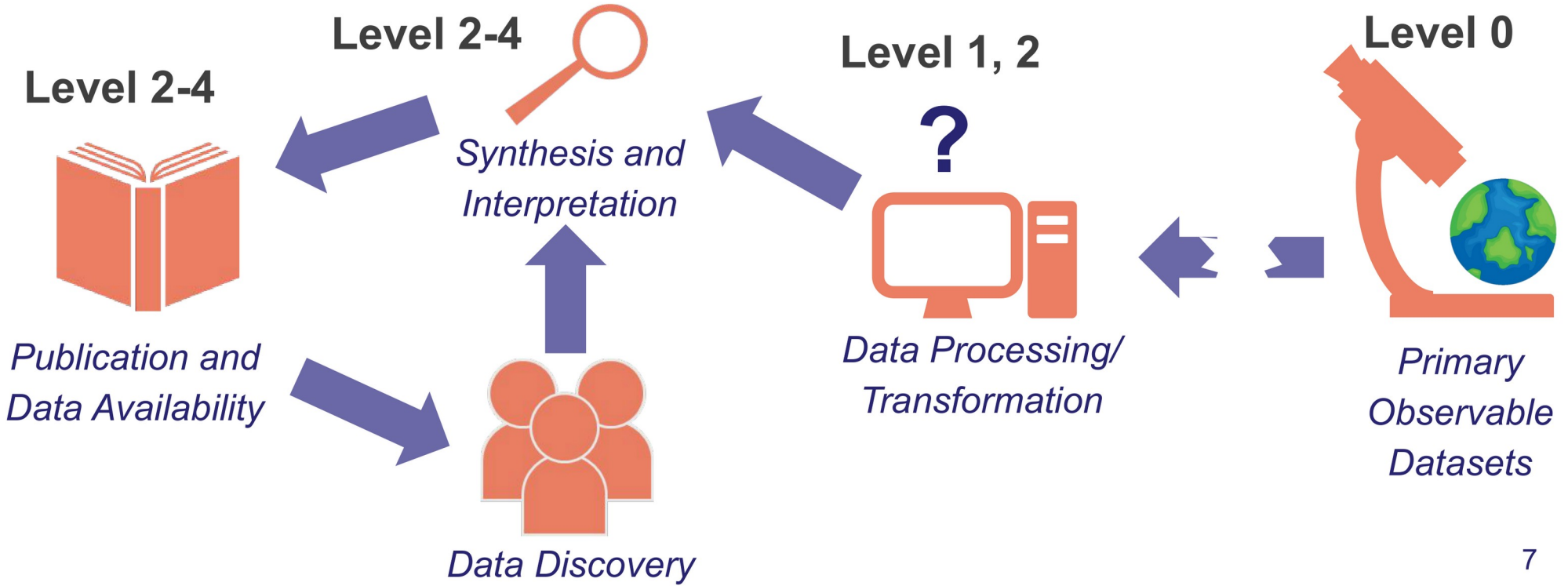
Gaps in Data Lifecycle

Primary data is not commonly preserved and/or made available as part of standard research and data lifecycles - how can we change this?

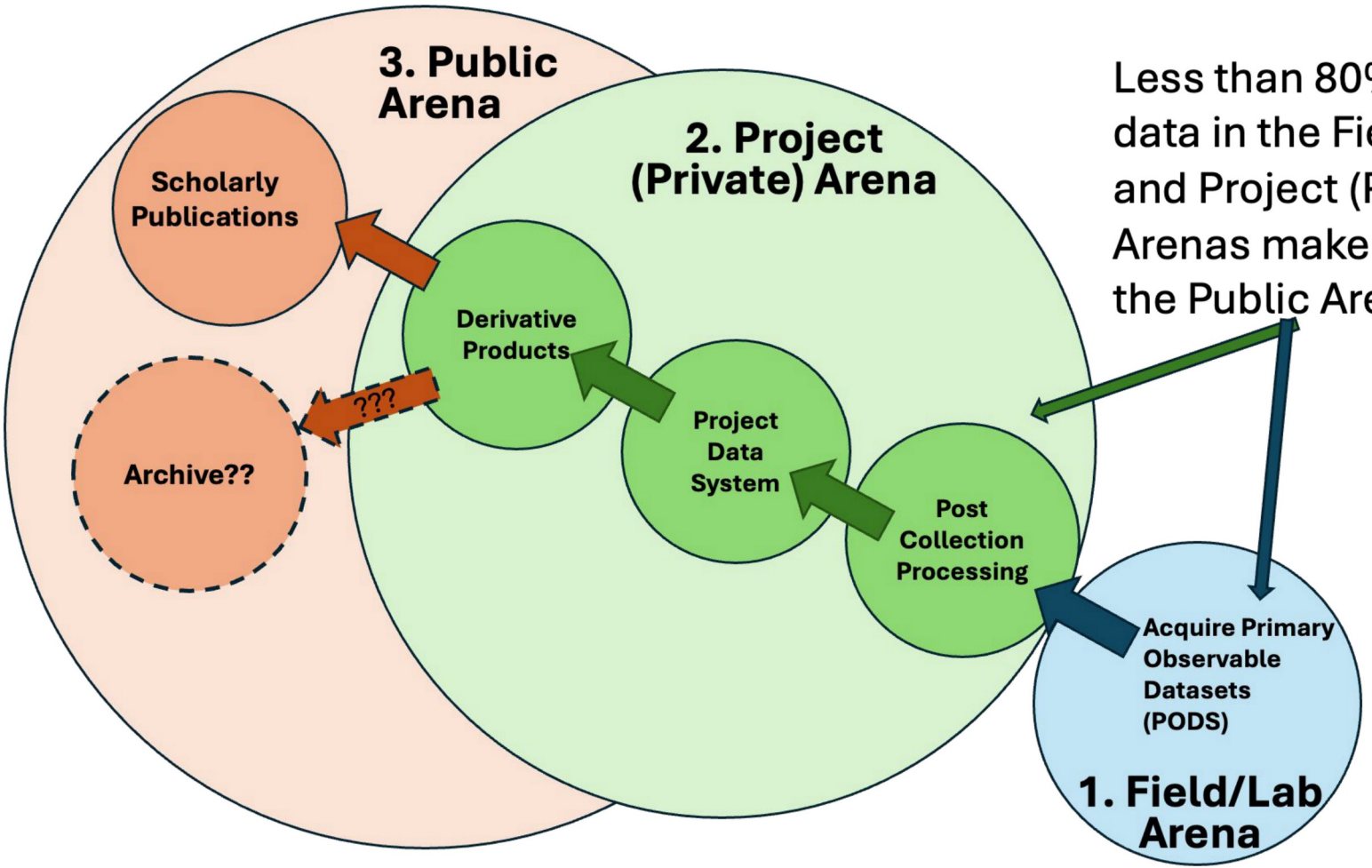


Gaps in Data Lifecycle

Primary data is not commonly preserved and/or made available as part of standard research and data lifecycles - how can we change this?



Traditional Reseach Methodologies are Linear and not Open

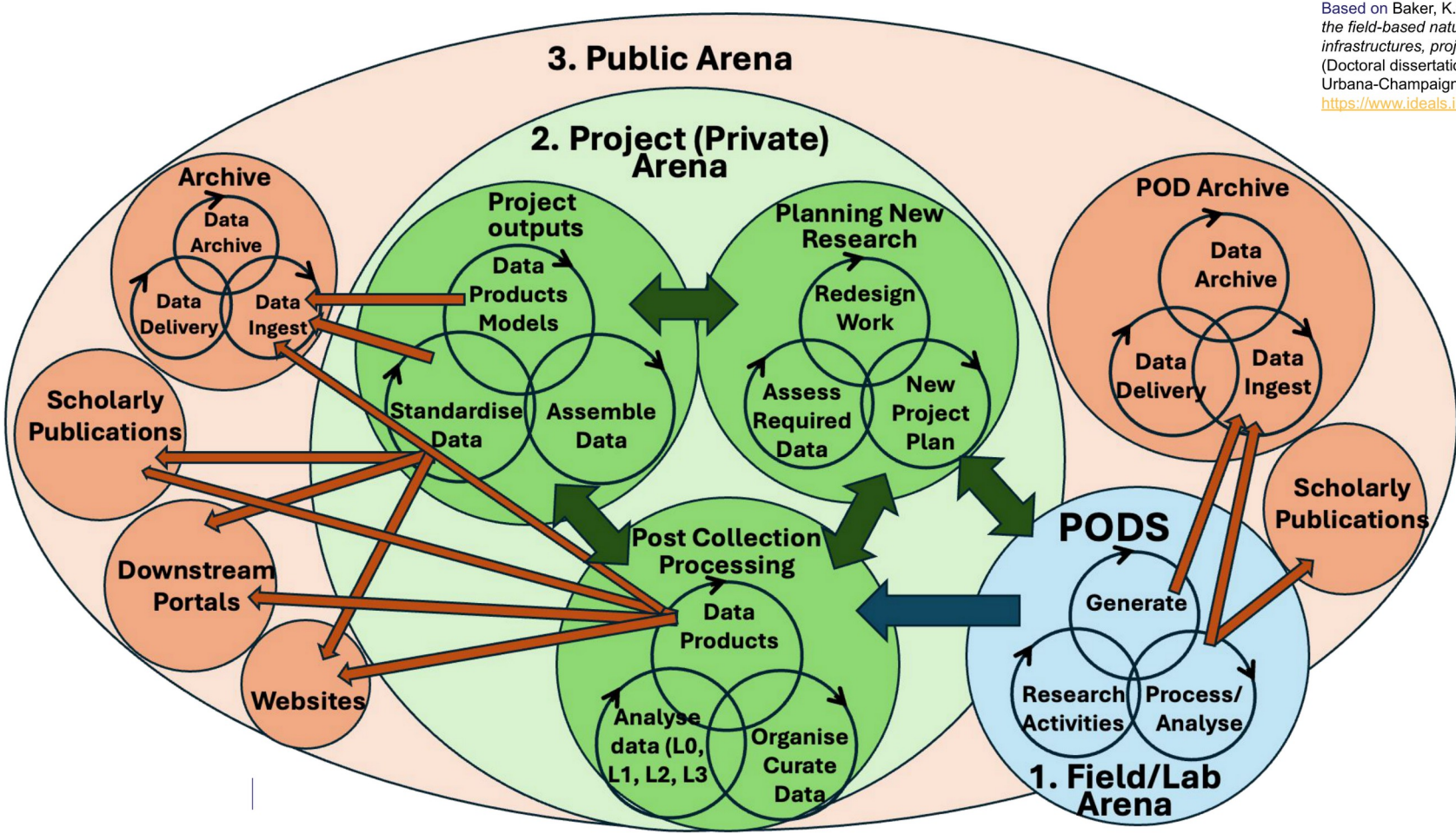


Less than 80% of data in the Field/Lab and Project (Private) Arenas make it to the Public Arena

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

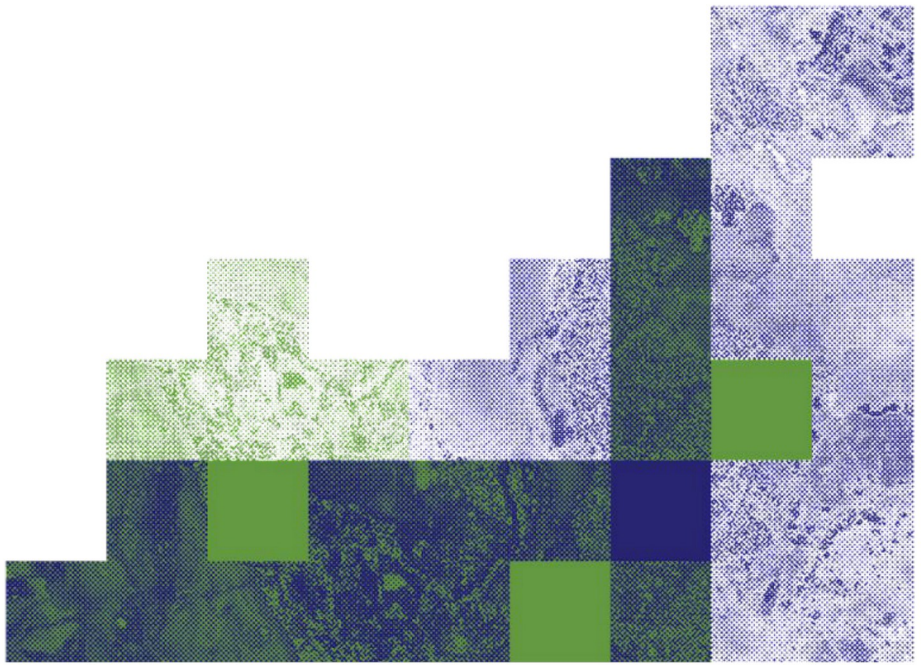
Based on Baker, K.S., 2017. *Data work configurations in the field-based natural sciences: mesoscale infrastructures, project collectives, and data gateways* (Doctoral dissertation, University of Illinois at Urbana-Champaign) (Figure 6.2, p 88) https://www.ideals.illinois.edu/items/103276



Cross Domain Interoperability Framework

Presenter:

Simon Hodson | CODATA



COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

PODs, POVs, and Metadata

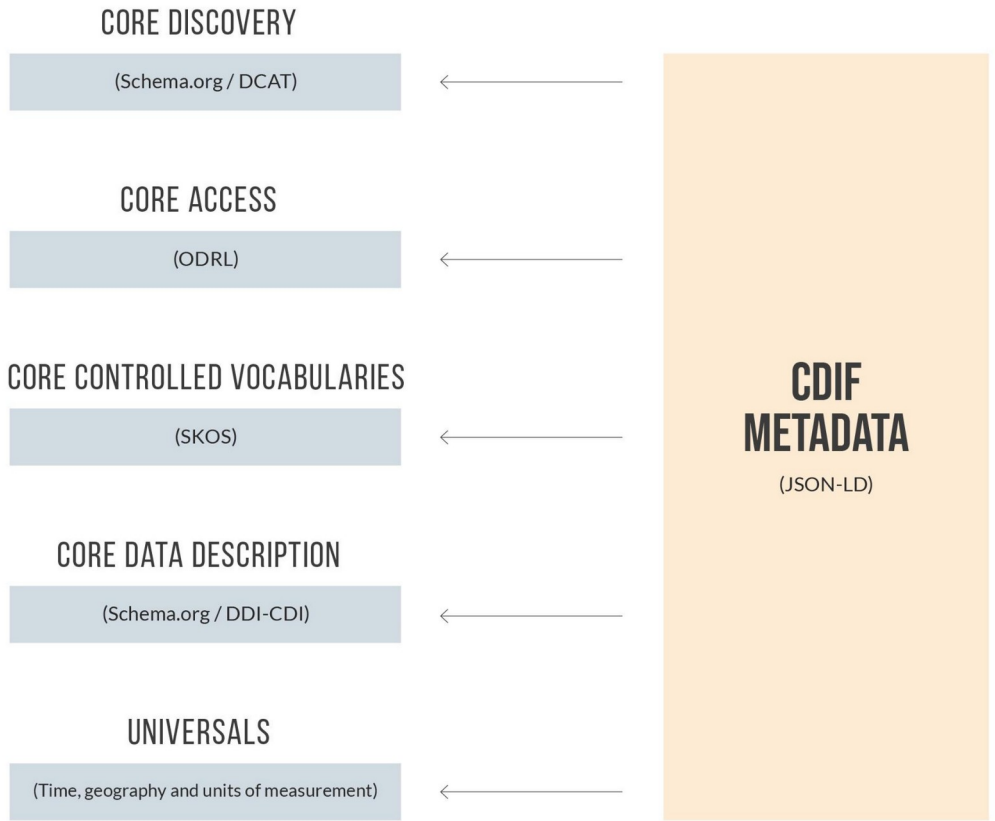
- PODs are fundamental to science, to principles of reproducibility and replicability.
- It is likely that there will be other future uses (aggregations, subsets) etc which may not be possible post processing.
- WorldFAIR Policy Brief ‘Enabling Global FAIR Data’: https://doi.org/10.5281/zenodo.11242702
  - Calls for a shift from a ‘bibliographic’ data stewardship practice to a data engineering practice.
- Return the observation and the variable to the heart of our thinking about data. Flipping the metadata model.

PODs, POVs, and Metadata

- What is the things being observed? ‘Object of interest’
- What is the observable property of that thing that acts as a proxy for the phenomenon we are interested in? ‘Conceptual variable’ (measurand, quantity… etc)
- What units are used in this measurement? Units of measure > cdi: representedVariable
- What is the estimate of error? What other quality-related information needs to be provided?
- What methods and techniques were used to create these measurements?
- CDIF work on context, provenance and quality… Intend to publish a first statement following Dagstuhl workshop in November.

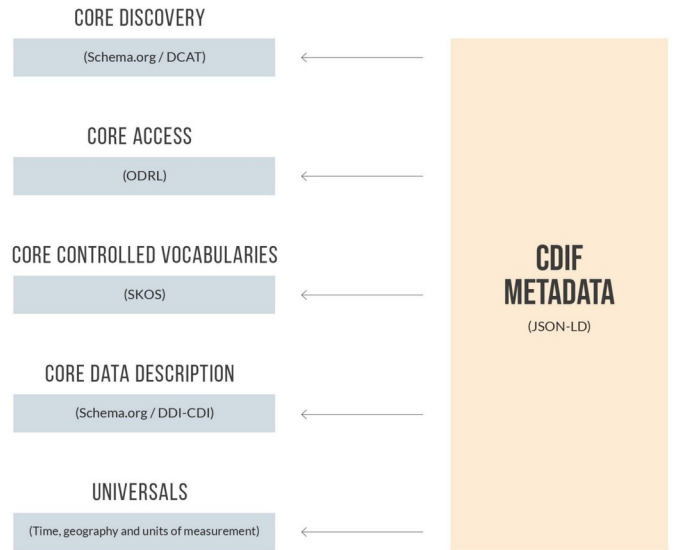
What is CDIF?

- The Cross Domain Interoperability Framework (CDIF) is a set of practical, implementation-level principles designed to improve data management practices within any community and lower the barriers to cross-domain data reuse. CDIF offers standards and methodologies for achieving different levels of interoperability necessary for reusing data across diverse domains. It is built around five core profiles that address the essential functions for implementing cross-domain FAIR principles.
- CDIF was first released in May 2024 as an output of the WorldFAIR project: https://doi.org/10.5281/zenodo.11236871
- The point of reference for CDIF and its component profiles is now the CDIF Book: https://cdif.codata.org
- **CDIF has attracted a lot of interest and has led directly to a set of additional projects and collaborations.**



Discovery Profile

- Discovery profile: https://bit.ly/cdif-discovery
  - A Content model that specifies the information expected to be included in any metadata record, with required, recommended and optional content items.
  - A JSON-LD serialization for that content using the Schema.org vocabulary to define the fields in a metadata record, and an implementation using the DCAT rdf vocabulary
  - Workflows to publish CDIF metadata so that is can be found and indexed by search providers using standard web technology
- **Variable description in the discovery metadata**
  - Name of the variable as it appears in the dataset.
    - Uses schema.org variableMeasured.
  - Text description.
  - propertyID with URI for the represented concept.

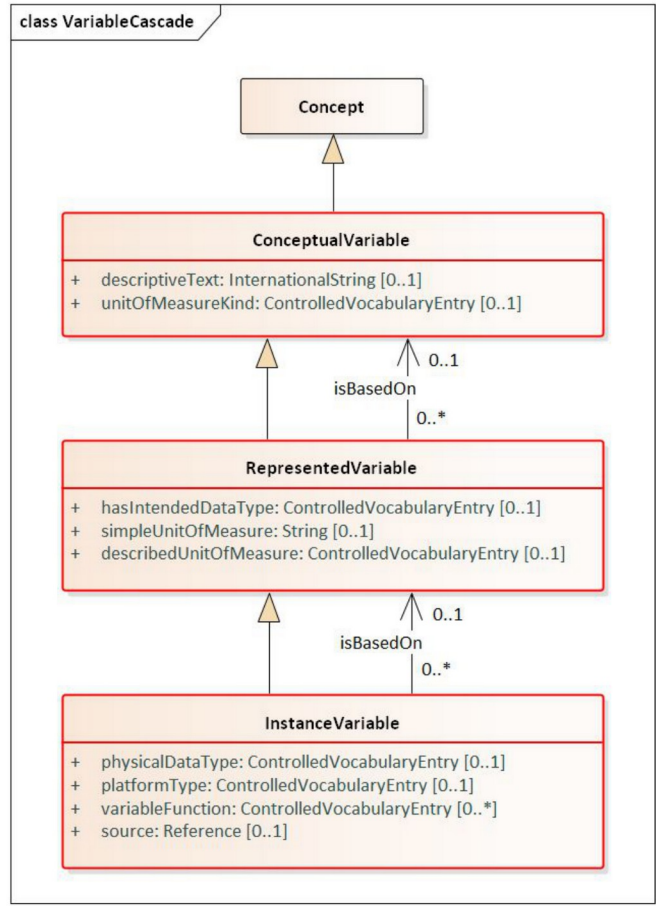


CDIF
Cross-Domain Interoperability Framework

Description Profile: DDI CDI for Data Structure, Variable Cascade, Provenance…

- Important to think about how we combine data for cross-domain research.
- Data Documentation Initiative (DDI) Cross-Domain Integration (CDI) specification contains three modules to assist with this:
  - **Structural Description:** assists processing of data structure transformations across four data structures.
  - **Data Description / Variable Cascade** describes data at an atomic level, describes relationships between concepts, representations and instances (assists with combining data and documenting information loss).
  - **Provenance and Processing:** module uses PROV-O and SDTL to provide and relay provenance and processing information.
  - Now officially released: https://ddialliance.org/ddi-cdi

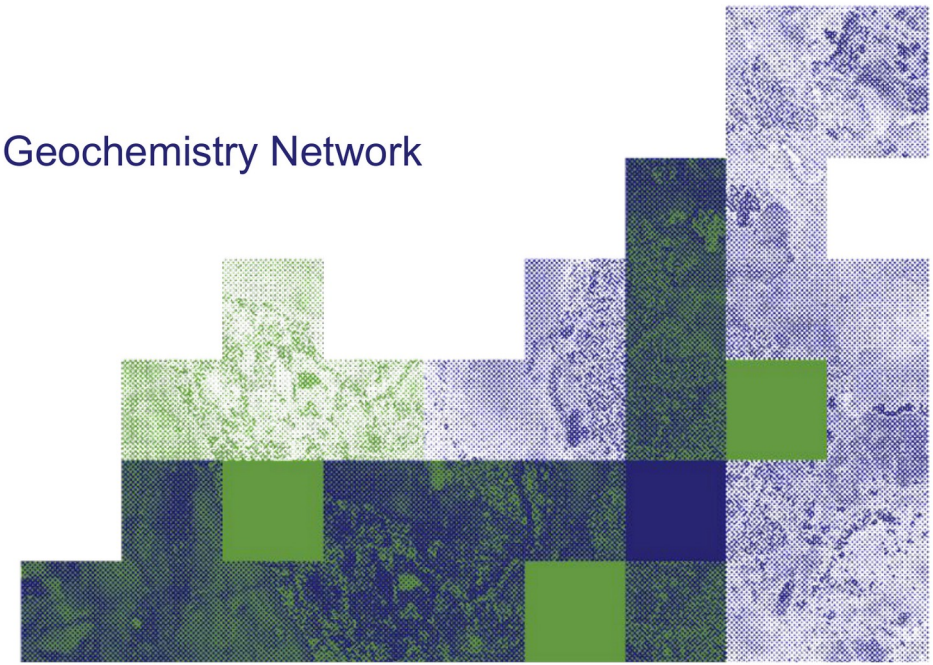
DATA DOCUMENTATION INITIATIVE



Case Study #1: Geochemistry

Presenter:

Angus Nixon | University of Adelaide | AuScope Geochemistry Network



THE UNIVERSITY of ADELAIDE

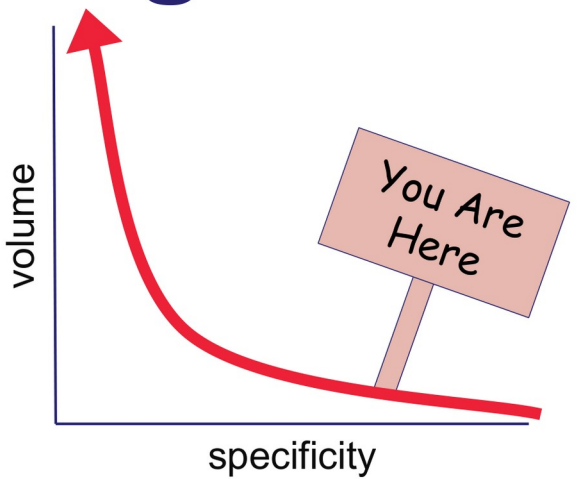
AuScope Geochemistry Network

AuScope

NCRIS National Research Infrastructure for Australia An Australian Government Initiative

Geochemistry - the ‘long tail’

- ‘Long-tail’ datasets instrument generated data
  - e.g. counts per second, voltage, imagery
- Low data volumes, high specificity



- No current systems to store PODs (?) - most primary data is stored on institutional drives or by individual researchers (hard drives, USBs, etc.)
- PODs accessibility allows data levelling, recalculation of ages/abundances with changing decay constants / reference materials, improving/iterating data modelling, data transparency

PODs for $^{40}Ar/^{39}Ar$ Geochronology

In 2025 AuScope awarded an opportunity fund to explore full lifecycle data storage and attribution of PODs for Ar-Ar geochronology

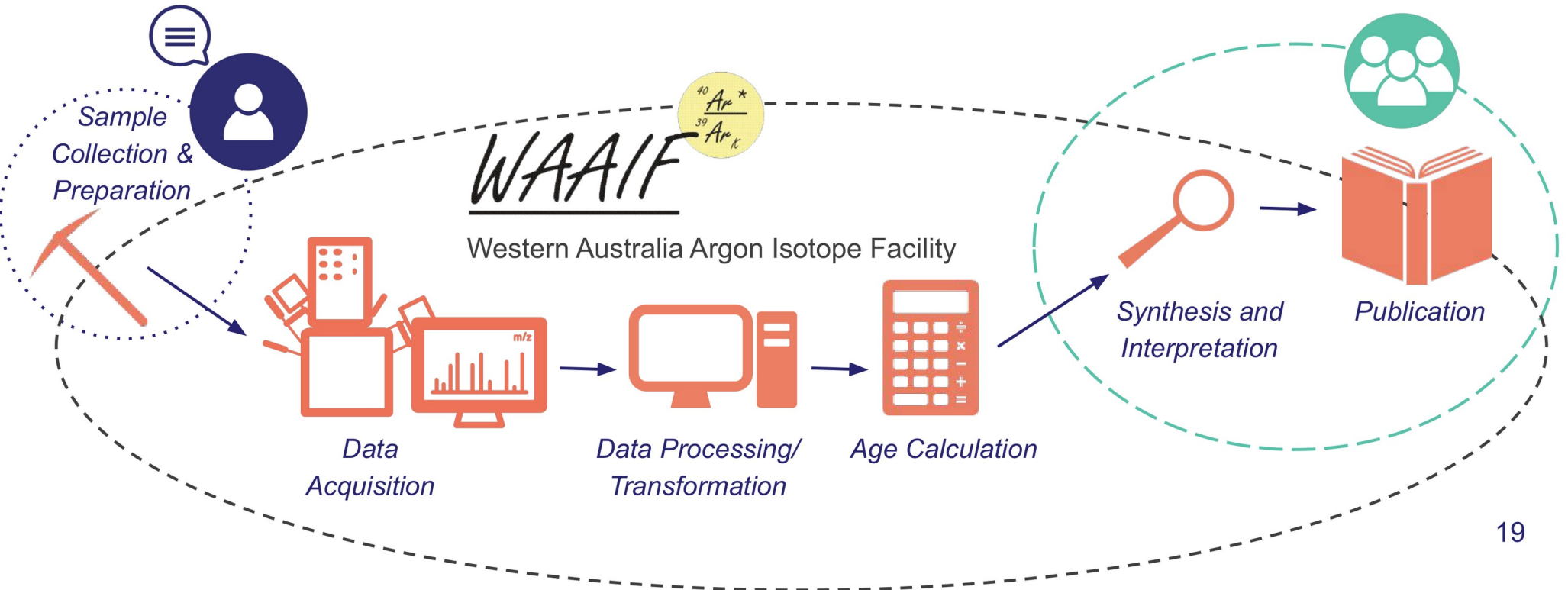
Seeds of Science - Establishing a Prototype Repository for Australian Primary Observable Data

AuScope

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

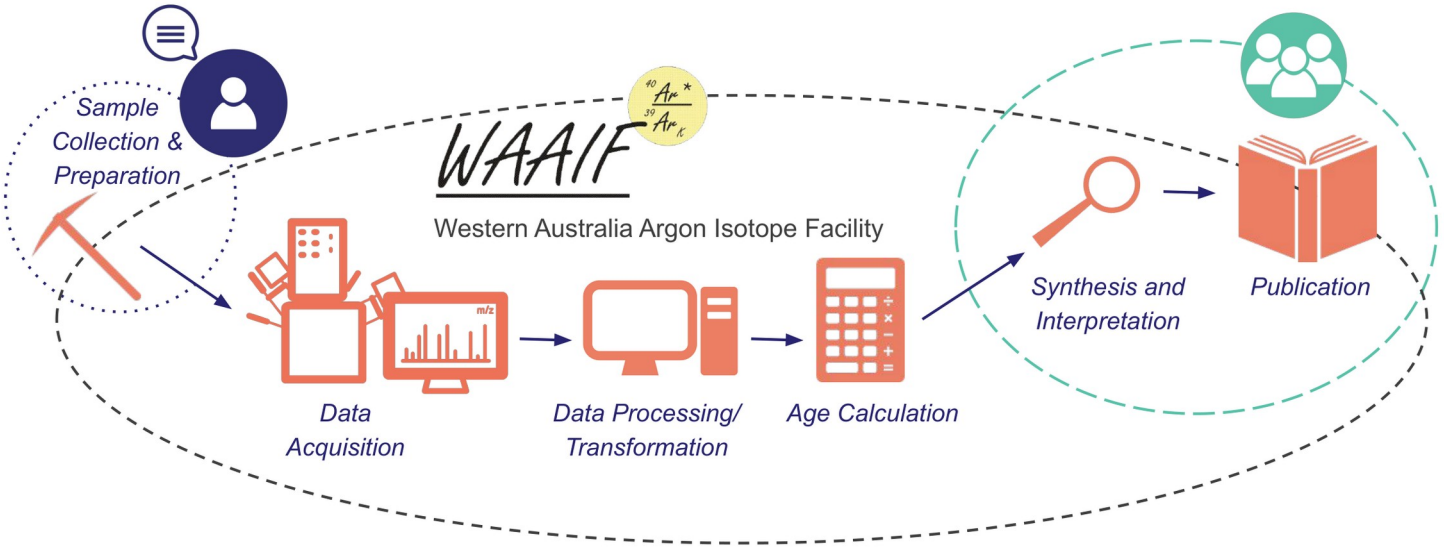
PODs for $^{40}Ar/^{39}Ar$ Geochronology

In 2025 AuScope awarded an opportunity fund to explore full lifecycle data storage and attribution of PODs for Ar-Ar geochronology - **why Ar-Ar Geochronology?**



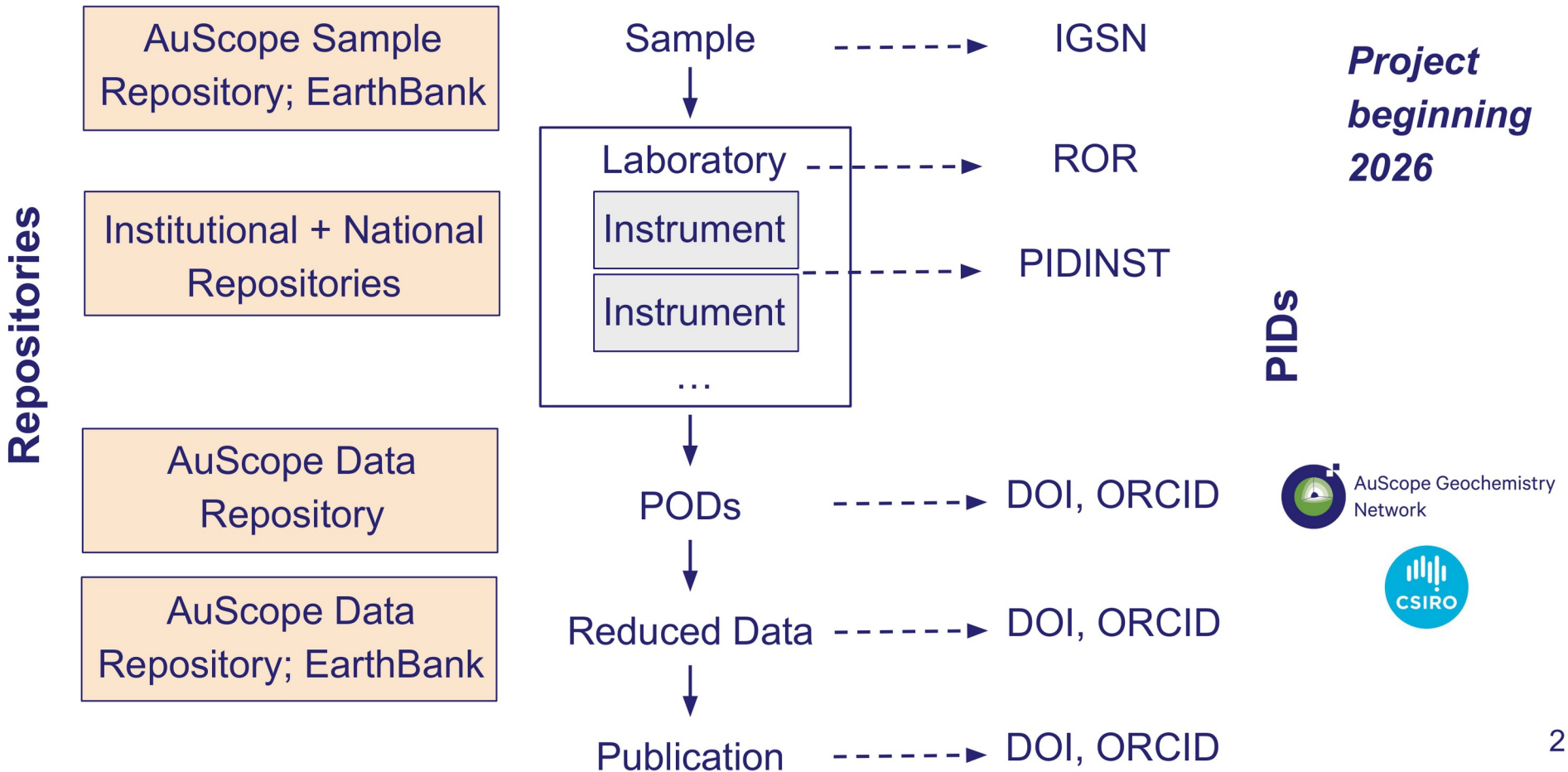
PODs for $^{40}Ar/^{39}Ar$ Geochronology

In 2025 AuScope awarded an opportunity fund to explore full lifecycle data storage and attribution of PODs for Ar-Ar geochronology



Framework for incorporating cascading PIDs, automated data harvesting, publication of PODs resources, linkage to research outputs

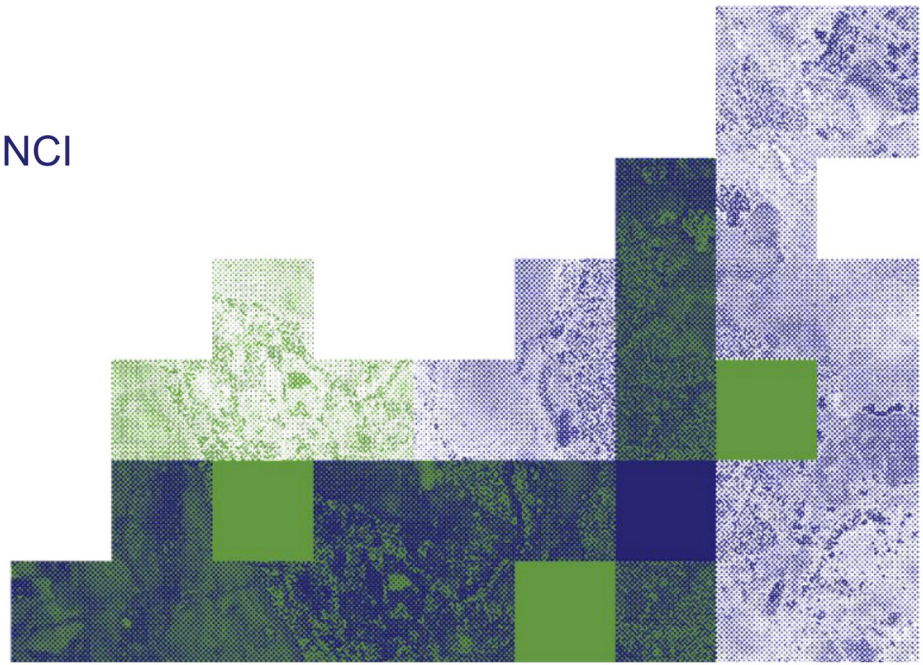
Digital Objects & Connections



Case Study #2: Geophysics

Presenter:

Lesley Wyborn | Australian National University | NCI



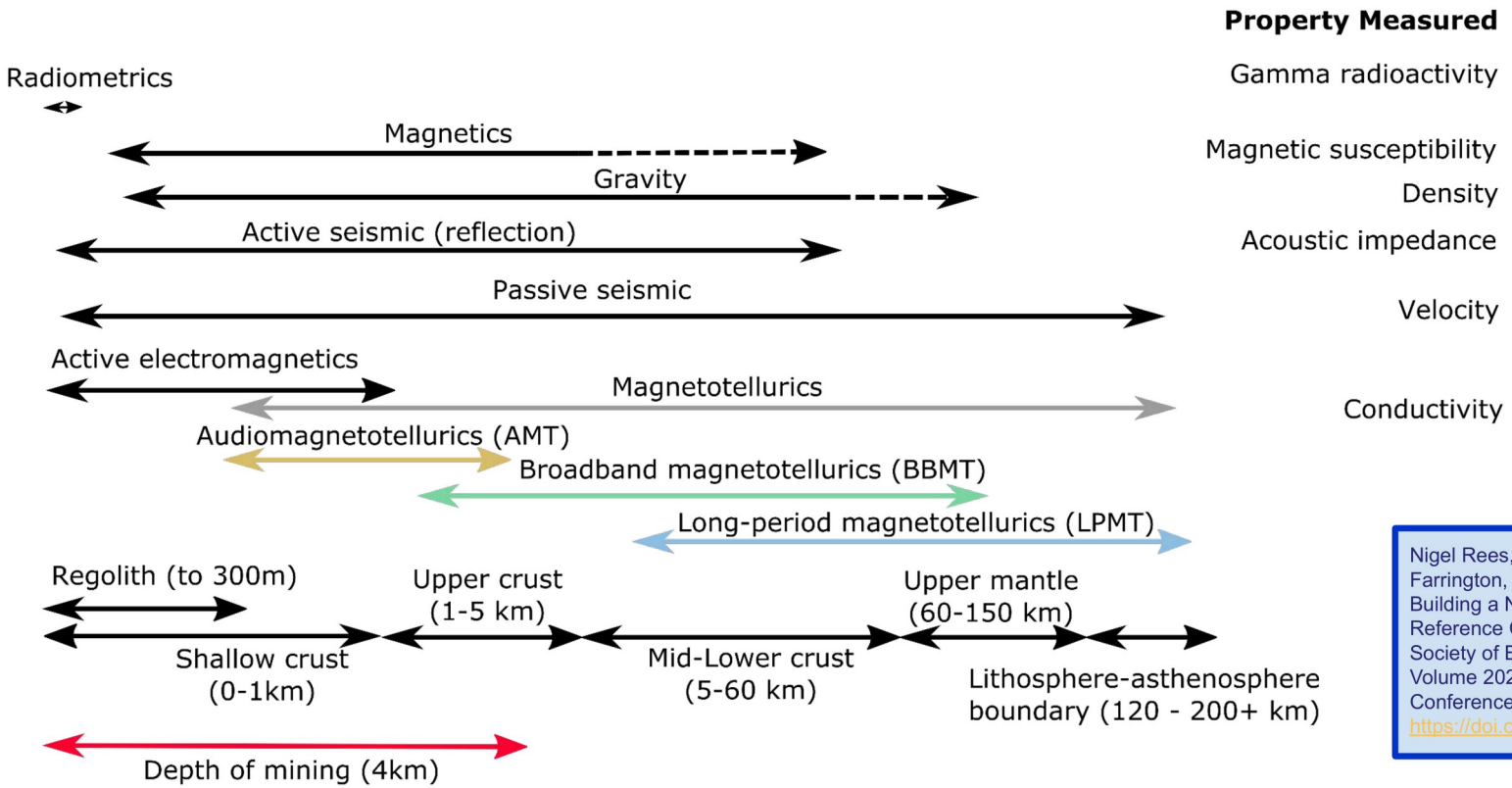
Australian National University

NCI AUSTRALIA

AuScope

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

Types of Geophysics High-res Data available at all crustal levels

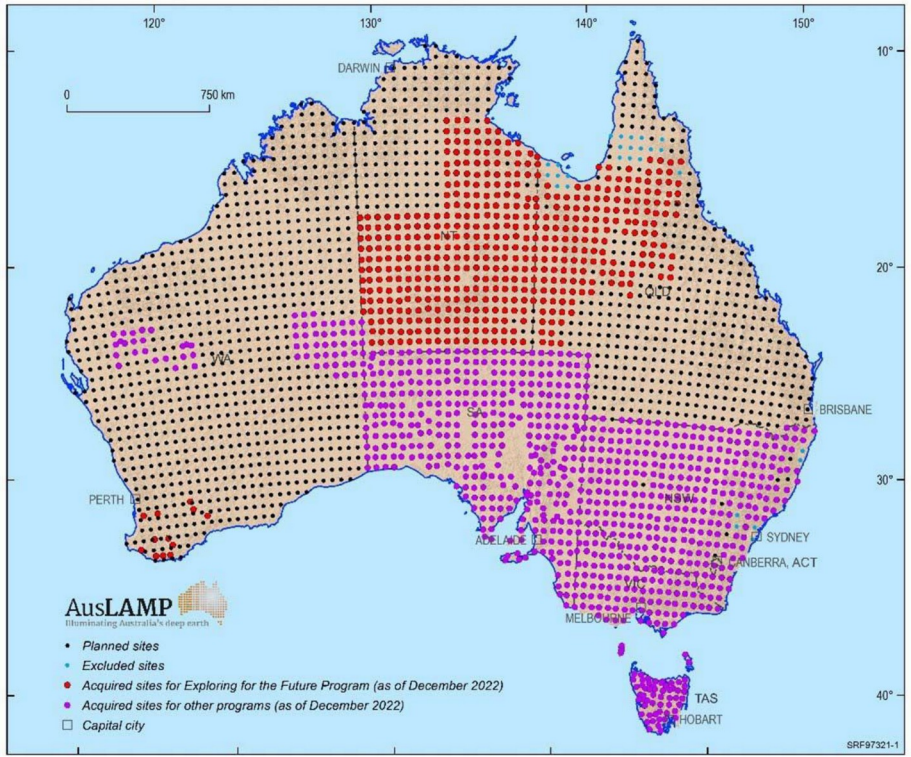


Nigel Rees, Lesley Wyborn, Ben Evans, Rebecca Farrington, Tim Rawling, Rui Yang, & Yue Sun. (2023). Building a National High-Resolution Geophysics Reference Collection for 2030 Computation. Australian Society of Exploration Geophysicists Extended Abstracts, Volume 2023, 4th Australasian Exploration Geoscience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

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).

lesley.wyborn@anu.edu.au

Example of National Scale Analysis: The Australian Lithospheric Architecture Magnetotelluric Project (AusLAMP) 2014-??



https://www.ga.gov.au/about/projects/resources/auslamp

- A collaborative project between Geoscience Australia, the state and Northern Territory geological surveys, AuScope, universities and other research organisations.
- Aims to acquire long-period MT data at approximately 3000 sites across Australia.
- Data are collected by multiple groups, using different instruments with very different infrastructure capacities, different formats, vocabs et.

lesley.wyborn@anu.edu.au

The basics of MT time series acquisition and processing

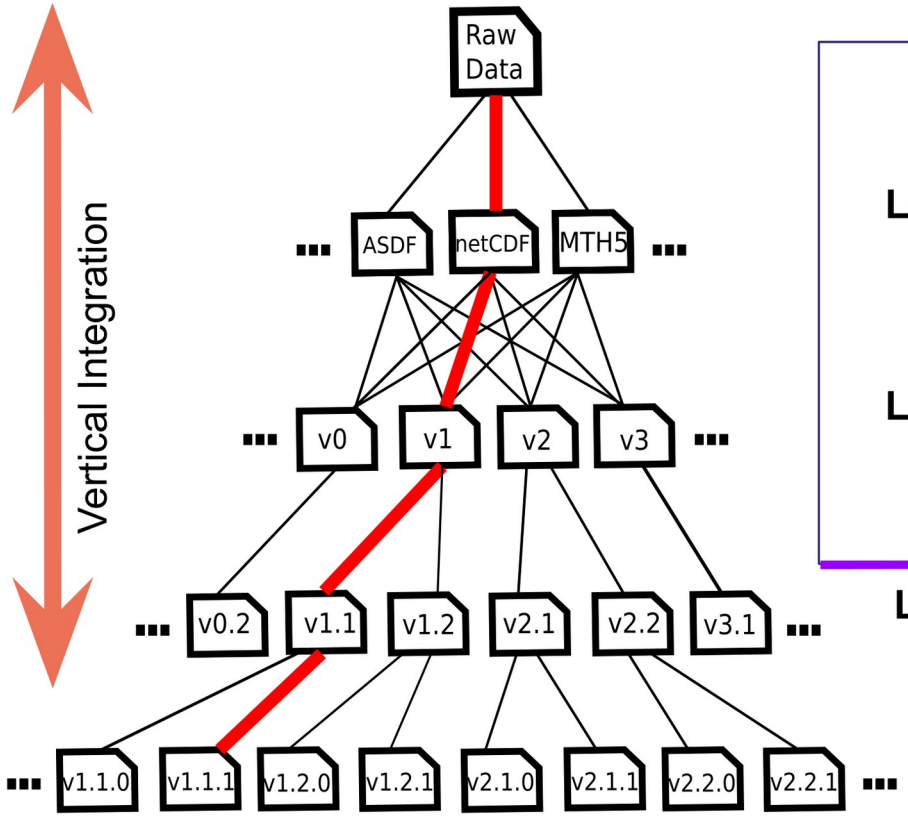
The **M**agneto**T**ellurics **t**ime **s**eries **d**ata **p**ublication (**MTtsdp**) codes: https://github.com/nci/MTtsdp

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

lesley.wyborn@anu.edu.au

For research innovation we needed access to all processing levels: Most PODS are currently invisible



Hard to get (Invisible?)

Level 0: Edited raw time series magnetotelluric (MT) survey data in different formats

Connected via PIDs

Level 1: Calibrated, resampled MT time series data

Connected via PIDs

GB-TB

Level 2: Derived frequency domain processed data (EDI Files)

Connected via PIDs

MB

Level 3: MT Modelling input/output files

lesley.wyborn@anu.edu.au

The great South Australian AusLAMP MT Time Series

https://www.cleanpng.com/png-internet-scavenger-hunt-ocean-grove-camp-meeting-a-1397955/

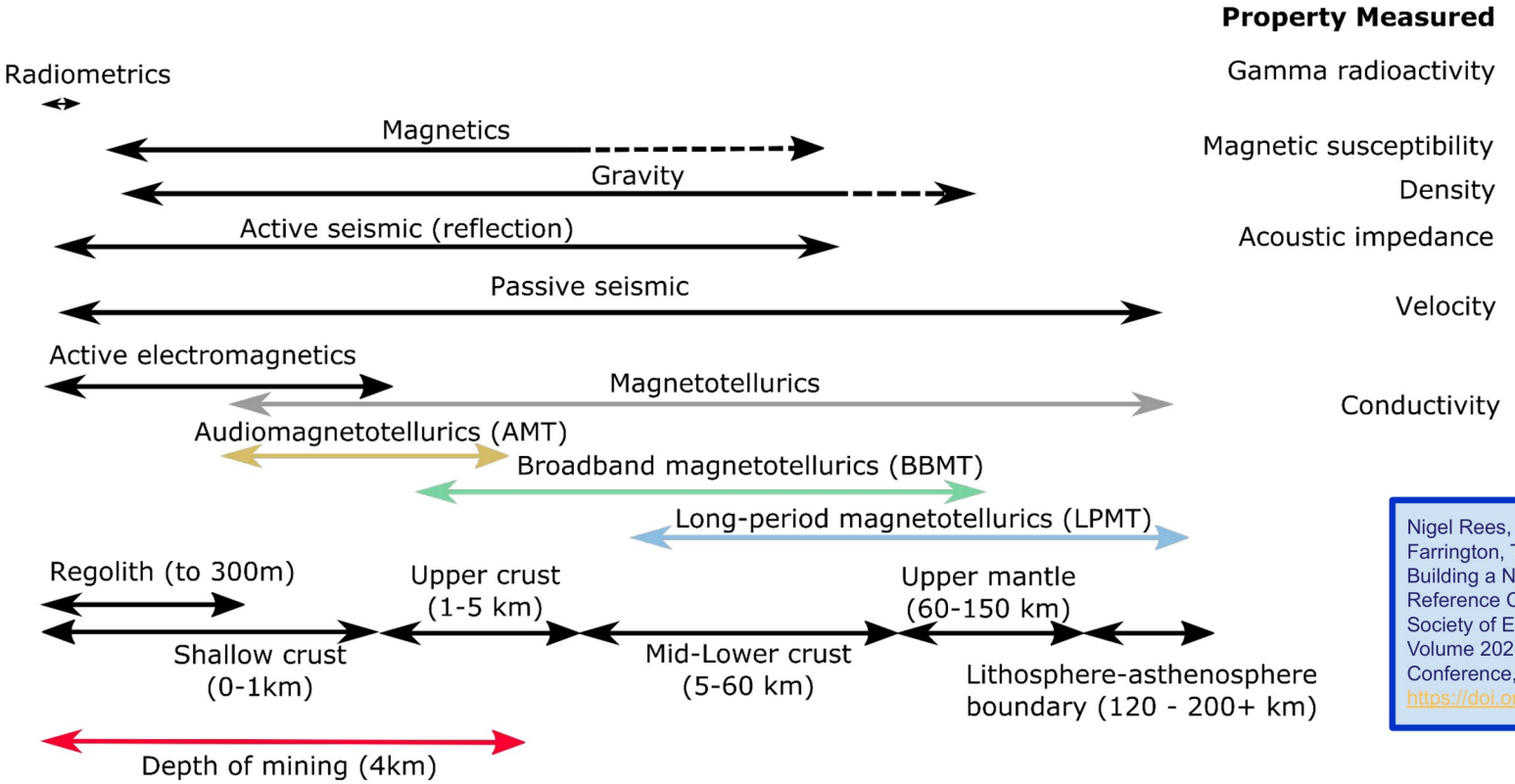
- **417** AusLAMP survey sites recorded in South Australia between 2014 and 2018: data were recorded on portable hard drives
- From 2020 to 2024, four comprehensive data rescue passes were conducted in order to maximise the number of sites recovered and ensure their availability at NCI.
- We recovered data on **400 sites** (we ran out of offices, basements, garages, etc, to search for hard drives, disks, old servers etc)
- We believe that HPC centres with their experience in managing highly processed data at tera- to petascales are ideally suited to store these large volume PODS, even if downsampled derivative products move to local repositories and cloud.

The great South Australian AusLAMP MT Time Series

https://www.cleanpng.com/png-internet-scavenger-hunt-ocean-grove-camp-meeting-a-1397955/

	Number of sites rescued	Percentage of total sites
Original ingest (2020):	205	49.1%
Second pass (December 2022):	270	64.7%
Third pass (June 2023):	386	92.5 %
Fourth pass (July 2023):	400	95.9%

Why we must preserve our PODS: High-res Multiphysics analysis at all crustal levels using the most modern techniques and algorithms



Bonsai-ed data is banned

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

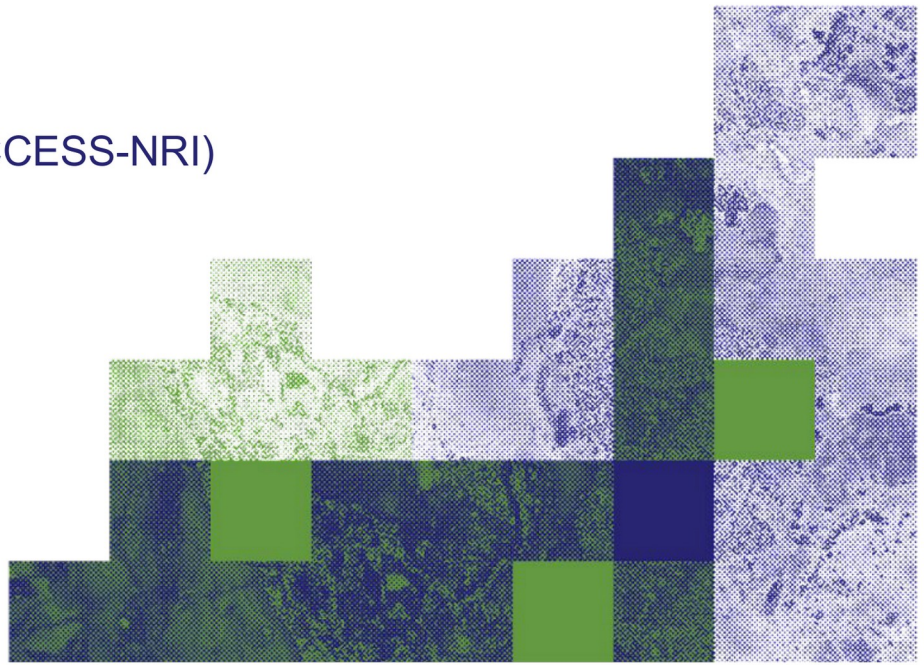
Types of geophysical data collected in Australia, the physical property measured and the depth of the crust that is sampled: also shown is the depth of current mining. Figure modified from original of Richard Chopping (GSWA).

lesley.wyborn@anu.edu.au

Case Study #3: Climate

Presenter:

Kelsey Druken | Australia's Climate Simulator (ACCESS-NRI)



ACCESS
National Research Infrastructure

Case Study #3 Climate

ACCESS National Research Infrastructure

Takeaways:

What types of data does your community work with?

- *Large netCDF4 simulation outputs (4-D)*

Are there current systems to store PODs? If so what are they, are they widely adopted?

- *Not really, equivalent 'native' output hard to work and not broadly usable.*

Why is it important to preserve PODs in your community?

- *Big drivers include cost (storage), reproducibility, and fit-for-purpose (lower-level versions more suitable for driving use-cases).*

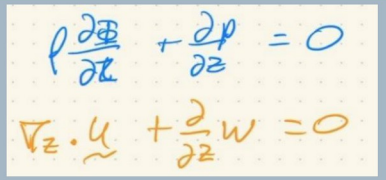
Australia's Climate Simulator (ACCESS-NRI)

ACCESS
National Research Infrastructure

- National research infrastructure funded by the Australian government (NCRIS) to support climate modelling (software and data)
- Provides computer simulations for climate, weather and Earth systems, specifically designed for Australia and the Southern Hemisphere.

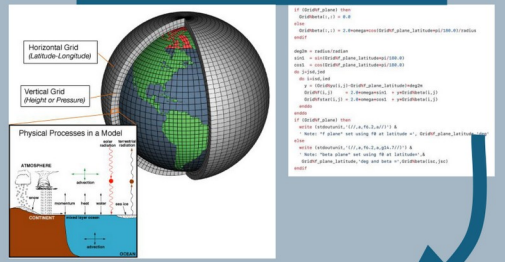
Software and data workflow

Equations based on fundamental physics

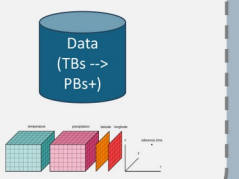


Supercomputer

Numerical simulations

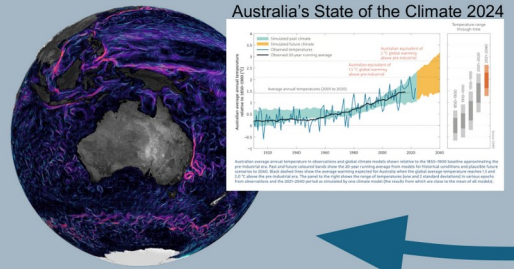


Model output data



Analysis code

Model visualisation & downstream climate/environmental data products



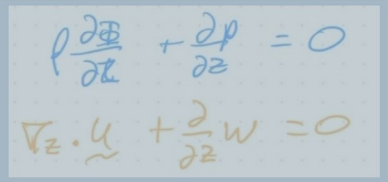
Australia's Climate Simulator (ACCESS-NRI)

ACCESS
National Research Infrastructure

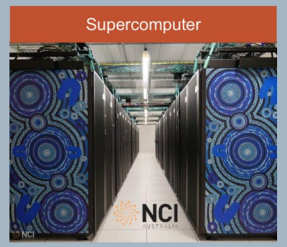
- National research infrastructure funded by the Australian government (NCRIS) to support climate modelling (software and data)
- Provides computer simulations for climate, weather and Earth systems, specifically designed for Australia and the Southern Hemisphere.

Software and data workflow

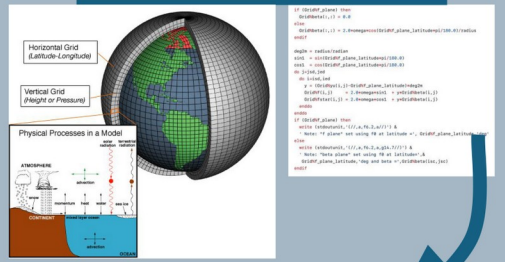
Equations based on fundamental physics



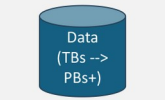
Supercomputer



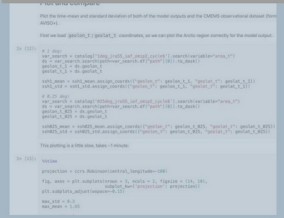
Numerical simulations



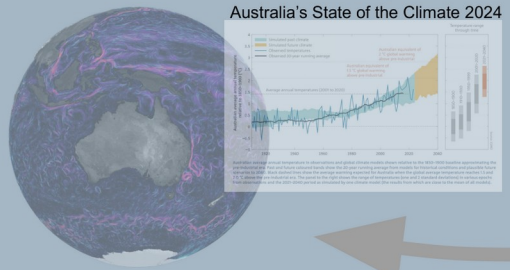
Model output data



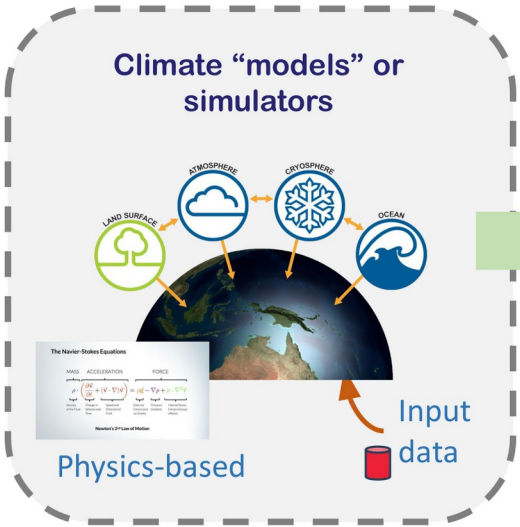
Analysis code

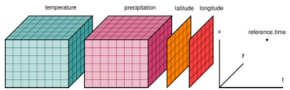


Model visualisation & downstream climate/environmental data products



Current state



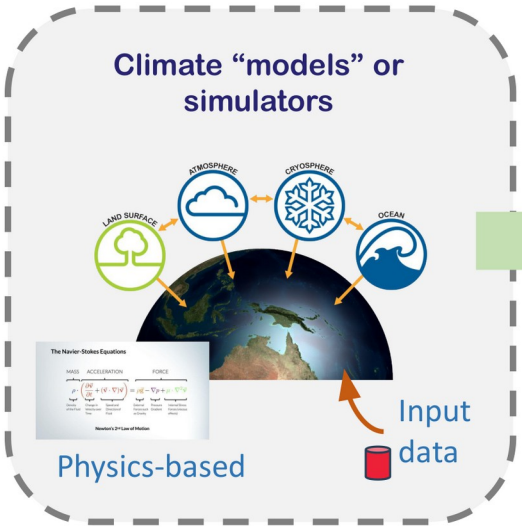


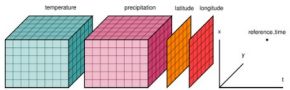
Native

Curation, standards and consistency maturity

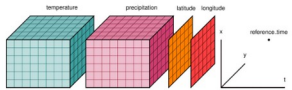
Time and effort

Current state





Native

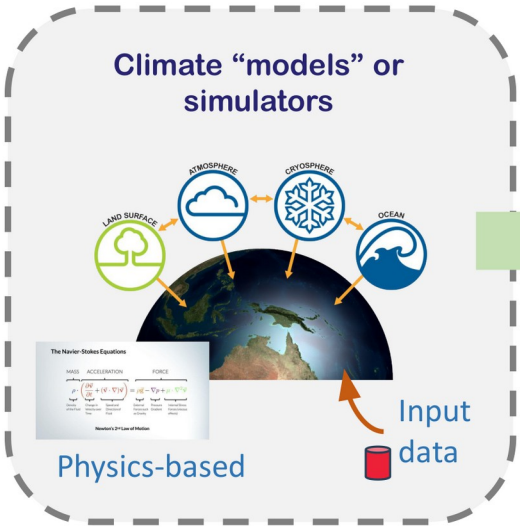


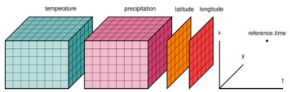
Processed (lightly)

Curation, standards and consistency maturity

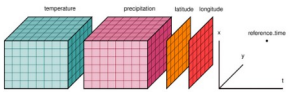
Time and effort

Current state

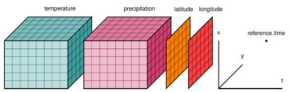




Native



Processed (lightly)



Processed (highly)

Curation, standards and consistency maturity

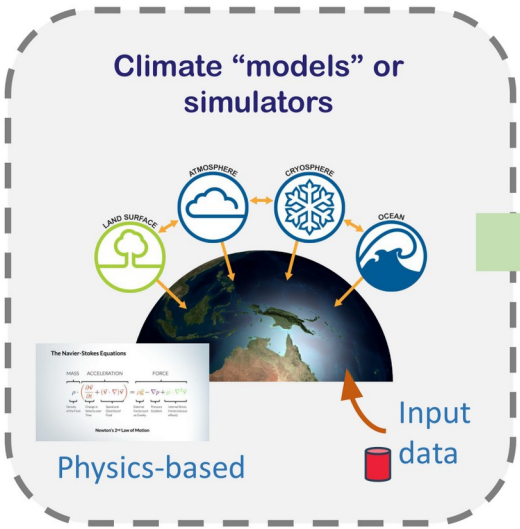
Time and effort

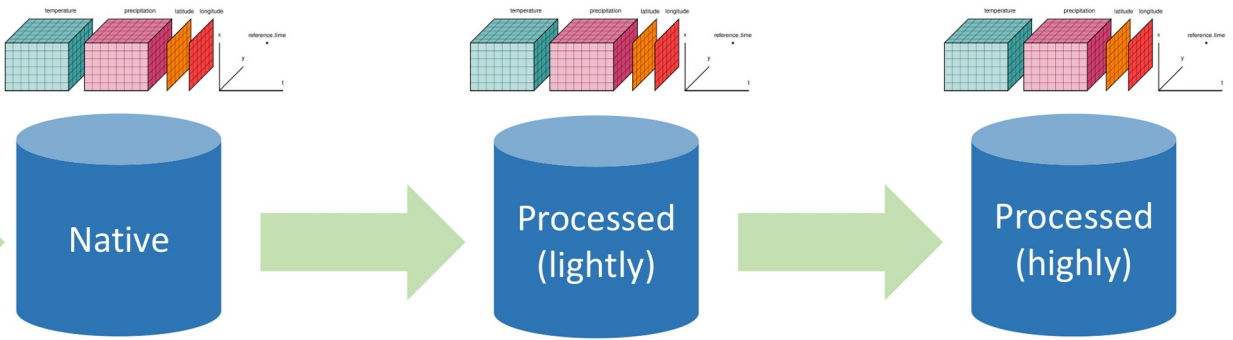
Community requirements

Use cases

#1. Model and data users want to use low-level data direct from the models

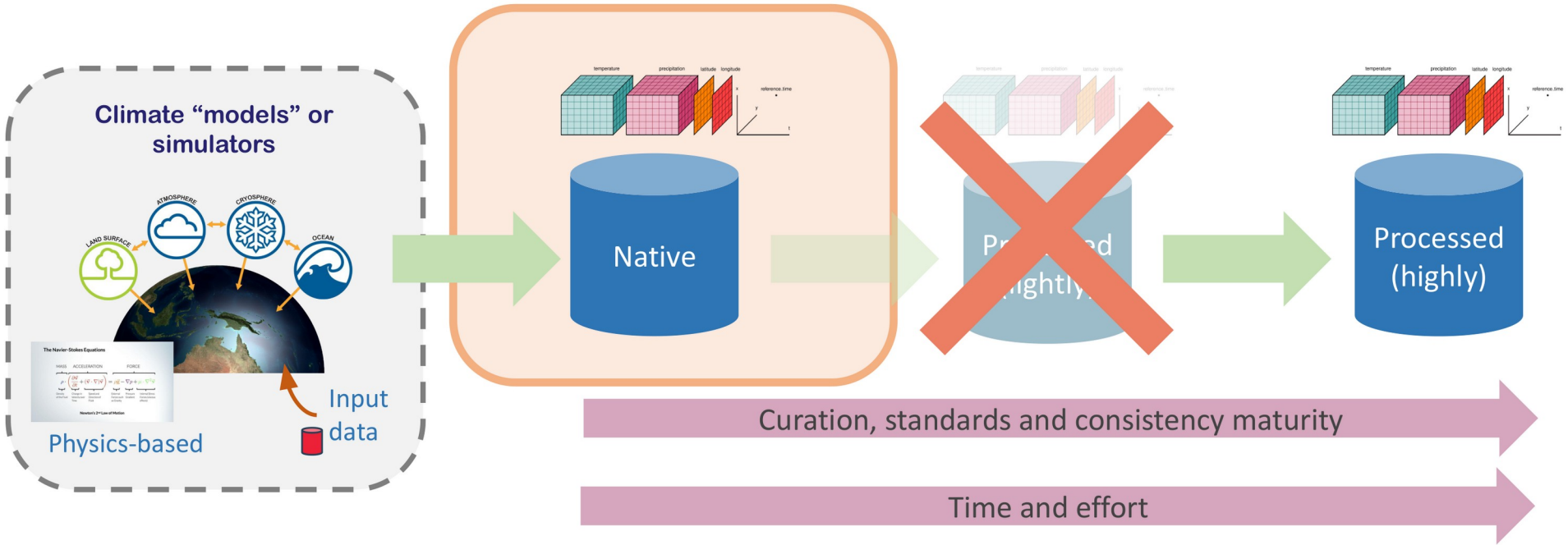
#2. ACCESS users want to contribute to or use outputs that comply with international model comparison specifications (e.g., CMIP7)





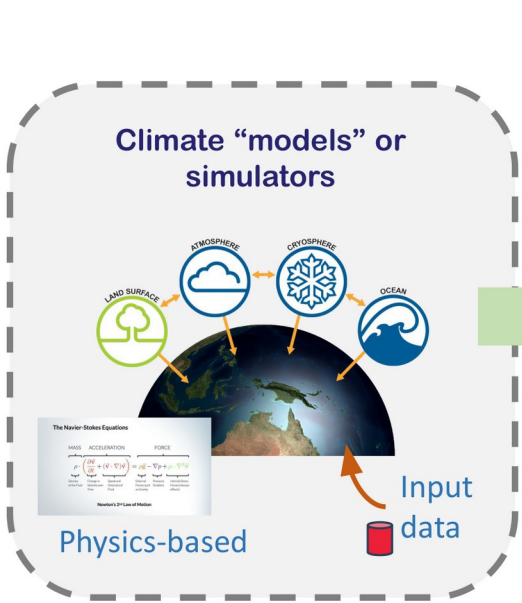
Storage: Each copy can range from 100s TBs up to PBs.

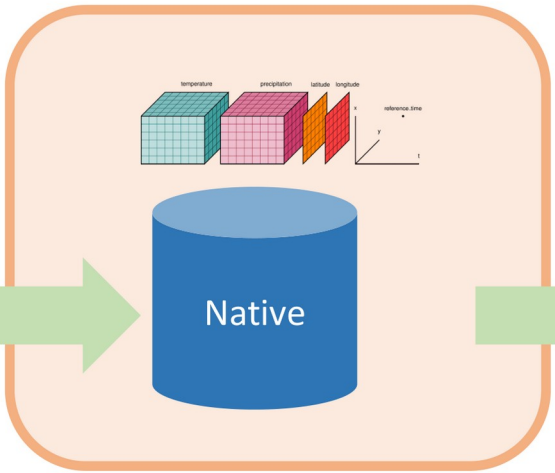
FAIR uplift – native outputs

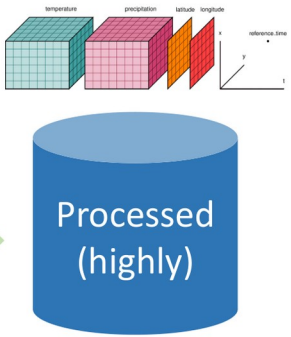


FAIR uplift – native outputs

Embed and automate data mgt from the start (not at the end!)



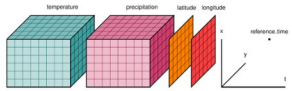




Curation, standards and consistency maturity

Time and effort

Embedding FAIR practices into the software



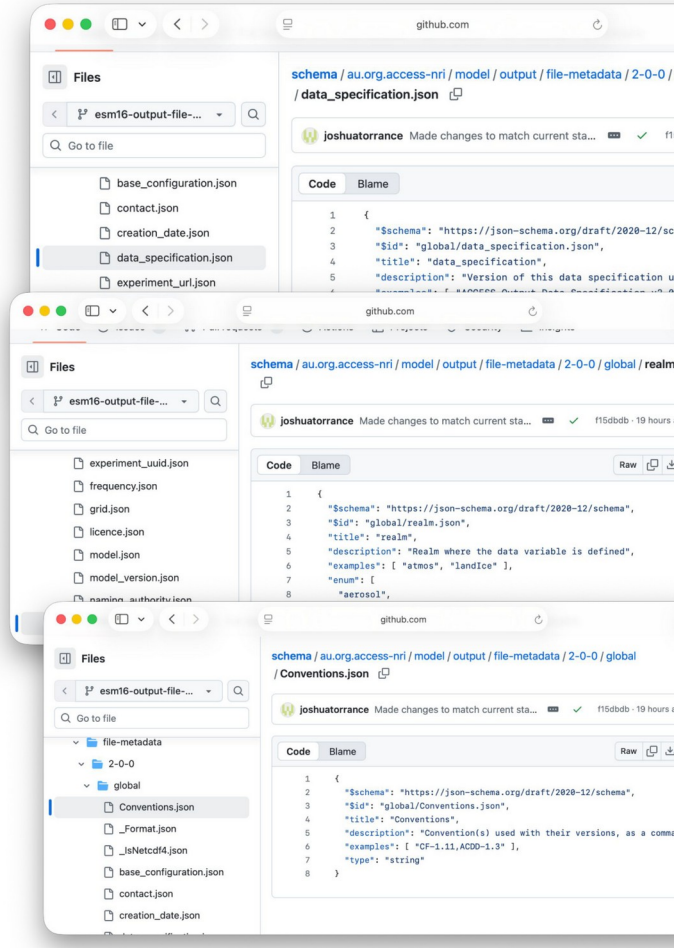
Outputs

Native (current)

- ❌ User documentation
- ❌ Naming convention
- ❌ International standards
- ❌ Controlled vocabularies
- ✅ Basic metadata
- ❌ Extended metadata
- ❌ Data specification

Native (future)

- ✅ User documentation
- ✅ Naming convention
- ✅ International standards
- ✅ Controlled vocabularies
- ✅ Basic metadata
- ✅ Extended metadata
- ✅ Data specification



Embedding FAIR practices into the software

ACCESS National Research Infrastructure

Native (cur

❌ User documer
❌ Naming conve
❌ International s
❌ Controlled vocabularies
✅ Basic metadat
❌ Extended met
❌ Data specifica

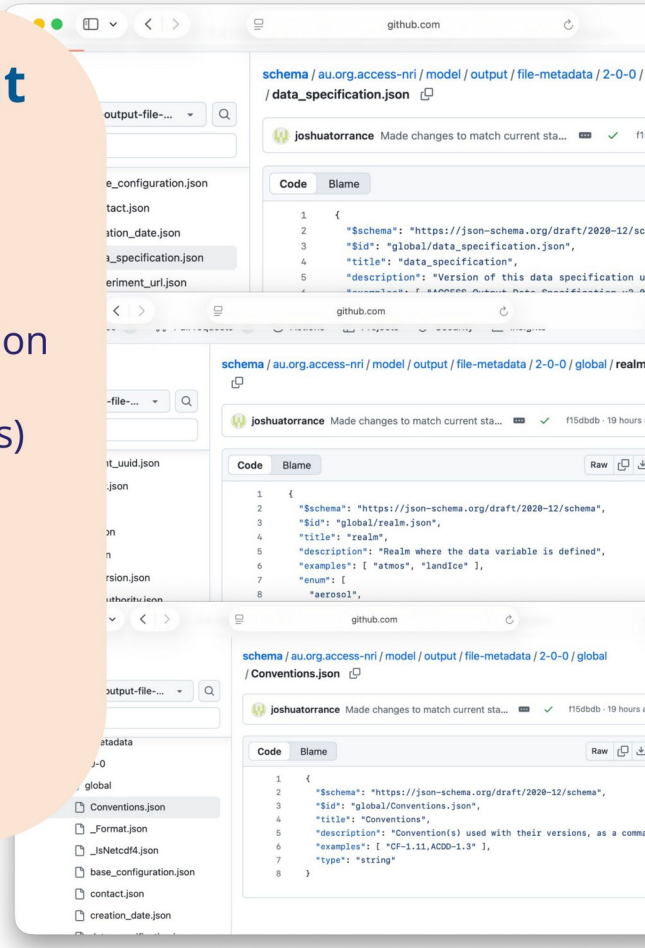
Drivers for uplifting our native output

→ **Storage is expensive**, hard to justify 3 different versions of the same datasets (when at PB-scale).

→ **We are human (we make mistakes)**, implementation of standards can be inconsistent across modelling components (e.g., atmosphere, ocean, and land models)

→ **Lowering barriers**, users currently want the lower-level data but need to have previous knowledge and understanding

→ **Improved support**, significant overheads for compatibility with data discovery and analysis tools



Discussion

- What kind of PODs does your community produce?
- Why might it be important to preserve these?
- Are there initiatives to preserve PODs in your community?
- What are the challenges for preserving PODs?
  - Technical
  - Social

Further Info

Dr. Angus Nixon

angus.nixon@adelaide.edu.au

Dr. Bryant Ware

bryant.ware@curtin.edu.au