SciDataCon 2025

Contribution ID: **137**

Type: **Session**

Building Earth and Environmental Science Data Repository Ecosystems: actioning locally - operationalising globally.

Tuesday 14 October 2025 11:30 (1h 30m)

Significance of the issues to be tackled:
Earth and environmental (E&E) datasets covering the six spheres of Earth System Science (geosphere, hydrosphere, biosphere, cryosphere, atmosphere, and anthroposphere) have been collected over centuries. Properly curated and preserved over time, E&E datasets can provide evidence-based inputs into longitudinal monitoring of changes over decades (e.g., desertification, sea level rise, anthropogenic contamination, climate variability, groundwater quality). When integrated with data from Health, Social Science and Humanities, E&E datasets make valuable contributions to the UN Sustainable Development Goals.

Early collections of E&E datasets were dominated by ground- or marine-based human observations and/or measurements by manual instruments, either in situ or laboratory-based. Airborne platforms emerged in the 1930s, followed by satellites (1960s), and Uncrewed Aerial Vehicles (UAVs, 2000s). Data acquisition underwent a paradigm shift in the 1950's with the computerisation of instruments, which quickly evolved to a capability for born-digital data outputs. Initially, most raw Primary Observational Datasets (PODs) were stored and managed locally within the institution that collected them; some PODs remained in this local state, others evolved into large-scale, internationally managed community resources. Additionally, non-observational datasets, including Climate/weather reanalysis and modelling, have emerged as fundamental datasets and are exposed to the same tensions as PODs.

Longitudinal requirements of E&E research necessitate that datasets be managed over decades and accommodate changes in instrumentation, hardware, standards, software, compute power, etc. Over time, the resolution and scale of PODs have increased: some data volumes from individual surveys are now in petabytes and require specialised HPC-D for storage and curation. Simultaneously, PODs collected as small-scale measurements are in megabytes and can be stored locally or on the cloud, yet their complexity and richness require specialised repositories to sustainably curate (meta)data to community-agreed domain standards. For both large and small volume PODs, the application of successive levels of processing throughout the data life cycle creates an incredible diversity of downstream datasets and products. In many cases, derivative products from very large volume datasets can be megabytes or smaller, no longer needing HPC-D for storage.

Apart from this diversity of data types and volume, there is a range in researcher capability, from highly skilled users who expertly use PODs and minimally processed reference datasets at any scale, to those who depend on pre-processed datasets to answer their research questions or to fit the needs of their software applications (and sometimes the capacity of the hardware and bandwidth they have access to).

While actioning systems locally or nationally within a single research community can be achieved, systems need to be operationalised within the global context and compatible with international strategies, including:

1. 2007 OECD Principles and Guidelines for Access to Research Data from Public Funding;
2. 2019 Beijing Declaration on Research Data: 'publicly funded research data should be interoperable, and preferably without further manipulation or conversion, to facilitate their broad reuse in scientific research';
3. 2021 UNESCO Recommendations on Open Science for open access to data, both raw and processed, and the accompanying metadata, analysis code and work flows;
4. FAIR Guiding Principles for Scientific Data Management and Stewardship;
5. TRUST Principles for Digital Repositories;

6. CARE Principles of Indigenous Data Governance.

It is nearly impossible for a single repository to meet all these requirements for every E&E data type and to serve all users. A 'Repository Ecosystem'is needed, one that balances and emphasises resources across the full data cycle and meets multiple considerations including:

1. Curation and sustainable preservation of raw full-resolution PODs captured directly off the instruments, under the expectation of limited need for ongoing access;
2. Calibration and conversion of the raw PODs into full-resolution reference datasets using community-agreed formats, data standards, etc. and their annotation with rich, FAIR- and CARE-compliant metadata. Many E&E PODs are dynamic and require regular updates with new data, calibrations, standards and technology. Once standardised, these high-resolution datasets can be aggregated into national/global datasets. These upstream datasets will mainly be accessed by power users;
3. Systematic reprocessing of full-resolution reference datasets into reusable downstream analysis-ready products, including mapping to uniform space-time grids, model outputs, syntheses and subsampling products. These can be accessed from distributed data platforms, virtual laboratories, portals, dashboards, etc. that are customised for specific user communities.

Separating the PODs'curation, preservation, calibration and conversion in 1 and 2 from the short-term and ever-changing distribution environments of analysis-ready products in 3, highlights the need for curation and preservation of raw full-resolution PODs to enable sustainable reuse over time, thus ensuring a capability to generate new knowledge in future scientific research.

But sustainable management of E&E data needs to also consider geopolitical changes that can abruptly impact repositories caring for key datasets. Securing these datasets for future use will depend on multiple countries proactively collaborating to ensure adequate redundancy and risk management.

Approach, structure, format, and suggested agenda:
This session will explore national actioning approaches for E&E Data Ecosystems that are operationalised globally around the above considerations.

1. Introduction (10 Minutes)
2. Lighting papers covering both national and international perspectives on linking E&E Data Ecosystems and Research Infrastructures (40 Minutes): Creating a National E&E distributed data ecosystem: catering for multiple users (Rebecca Farrington, AuScope); Global Ecosystem Research Infrastructures (Beryl Morris); Earth Science Research Infrastructures (Tim Rawling, AuScope); Oceans Data Information System (Speaker TBC); GBIF (Peggy Newman, Atlas of Living Australia); Connecting national E&E datasets to the WorldFAIR cross-domain project (Speaker TBC).
3. Community forum and determining the next steps (40 Minutes).

Primary authors: WYBORN, Lesley (Australian National University); FARRINGTON, Rebecca (AuScope); DRUKEN, Kelsey (ACCESS-NRI); HOBERN, Donald (Australian Plant Phenomics Facility); LESCINSKY, David (Geoscience Australia); NEWMAN, Peggy (Atlas of Living Australia); ROBINSON, Andrew (Australian National Computational Infrastructure)

Presenters: RAWLING, Tim (AuScope); WYBORN, Lesley (Australian National University); FARRINGTON, Rebecca (AuScope); NEWMAN, Peggy (Atlas of Living Australia); MORRIS, Beryl (TERN)

Track Classification: SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global