SciDataCon [part of IDW 2025]

Collaborative Notes

Session title	**Emerging technologies in the global context: Challenges and opportunities for the long-term environmental data management lifecycle**
Session chairs	Gretchen Stahlman (gs23j@fsu.edu) and Alison Specht (a.specht@uq.edu.au)
Session link for more information	https://scidatacon.org/event/9/contributions/35/
Session programme time	Tuesday, 14 October 23:00 – 00:30 UTC / Thursday, 16 October 11:30-13:00 AEST, room M4
slides	Bird, S., Hugo, W., Guru, S., Maurer, G.E., Stahlman, G., Stall, S., Su'a, T., Specht, A., 2025. Emerging technologies in the global context: Challenges and opportunities for the long-term environmental data management lifecycle. https://doi.org/10.5281/zenodo.17460604
Recording (available only to registrants until January 2026)	https://www.gevme.com/page/on-demand-ybdsx

Registrants:

Please complete this table to indicate your attendance (add rows as needed):

Name (first, last)	**organisation**	**country**	**email**	**Are you interested in future collaboration? (Y/N)**
Graham Parton	CEDA	UK	graham.parton@stfc.ac.uk	y
Shelley Stall	AGU	US	sstall@agu.org	Yes
Mingfang Wu	ARDC	Australia	mingfang.wu@ardc.edu.au	
Doug Schuster	NCAR	US	schuster@ucar.e	

			du	
Katie Hannan	CSIRO	Australia	katie.hannan@csiro.au	Yes
Richard Ferrers	ARDC	Australia	richard.ferrers@ardc.edu.au	
Brian Minihan	ORCID	US	b.minihan@orcid.org	

Session aims

This 90-minute session attempted to identify critical factors affecting the data lifecycle over the long term. These included the inherent expansion of data over time across domains and knowledge contexts, the effect of emerging technologies, the increasing energy and financial cost of handling, curating, and using data, and who takes responsibility for the curation of data for the common good. Environmental research data was our domain focus.
Invited panellists from various stages of the data lifecycle (see details below) used their experiences and how they have overcome the challenges of maintaining good practice against adversity to answer three questions. Each panellist presented pre-prepared short statements (not all panellists provided comments against every question) followed by a short Q & A with the audience.

The session was planned by a team led by Gretchen Stahlman (Florida State Uni, USA) and Alison Specht (TERN, UQ), and advised by Siddeswara Guru (TERN UQ), Inna Kouper (Indiana Uni, USA), Greg Maurer (New Mexico State Uni, USA), and Shelley Stall (American Geophysical Union).

This session was in panel format, for which we assembled seven experts (listed below) to consider three questions. The panellists selected questions to which they felt they could contribute best.

1. How do we balance the promise of emerging technologies with the practical risks of data loss and preservation challenges across the long-term data lifecycle?

Panellists: Stephen Bird, Wim Hugo, Gretchen Stahlman, Tavita Su'a

2. What does true democratization of environmental data look like—and who might be left out?

Panellists: Stephen Bird, Siddeswara Guru, Greg Maurer

3. How can we ensure that our environmental data management practices remain sustainable—both environmentally and ethically—in a rapidly shifting global context?

Panellists: Siddeswara Guru, Wim Hugo, Greg Maurer, Shelley Stall, Tavita Su'a

Meeting notes

Several of the panellists contributed written material for discussion which are recorded in the shared slides: https://doi.org/10.5281/zenodo.17460604

Others presented their thoughts and this is captured in the session recording.

Miro boards were used to provide a platform for audience members to contribute to each of the questions. The results of the Miro interaction are reported below.

Open comments on each discussion question:

Q1: How do we balance the promise of emerging technologies with the practical risks of data loss and preservation challenges across the long-term data lifecycle?

- Employ structured policies on what makes a dataset eligible to include a DOI.
- bot usage - what's legit (in terms of use) and what's just 'junk' usage? How do we distinguish 'genuine' use via 'bots' for allowing access/ reporting stats? Should we distinguish genuine 'bot' usage from human usage?
- Benefit can be contextual, so something useful in one context may not be useful in another context
- while technology accelerates in terms of scale and scope, how do we ensure we're maintaining data and resource accessibility as widely as possible?
- store data on multiple types of media, track data in non-proprietary mediums (e.g CSV dumps from relational databases and XML representations of metadata)
- Shares by Wim: PIDs don't guarantee access to data. After 12 years, only 50% are likely to resolve.
- Include policies for purging simulation outputs (not observations) once they are superseded by new products or no longer relevant
- store data on appropriately types of storage (e.g. Raw, unprocessed data on cold storage, popular variables from "Analysis Ready" datasets on high performance storage)
- an example from the UK is that we're having to consider repacking data to ensure that they can be accessed via the web, which wasn't picked up originally as principle use of the data was from local disk access. That meant there were external users who couldn't access the resource and so were silently going away.

Q2: What does true democratisation of environmental data look like—and who might be left out?

- Google Earth Engine has democratized access to many environmental datasets to under resourced communities.
- Create geographically distributed "convenience copies" of high value datasets.
- Responses (post sticky notes anywhere:
- **see data democracy essay at regional AU university - https://www.cerdi.edu.au/cb_pages/data_democracy_essay.php**
- how do we balance the need for democratisation around data decisions vs. timeliness of reaction that may be needed ? (e.g. in response to emerging needs on short timescales)
- How do we consider the 'end-of-life' for data too in a truly democratic way when we're faced with limited resources and the need to be netZero mindful?
- Builds on top of tech, network and skills, support, and sustainability

- All researchers have equitable access to compute co-located with data. (Not bandwidth limited)
- Data analysis platforms co-located with data repositories provide capabilities for a wide range of users from entry level to very experienced
- Researchers have an equitable repository resource to deposit and publish their research datasets. This includes data curation consulting support
- Consider applying CC BY-NC licenses to prevent for-profit commercial entities from overwhelming data repository capacity.
- Store data in non-proprietary, community accepted formats.
- People without access to technology are left out.
- From a group who produce long term monitoring products, we are limited by funding and human resources for what we can publish, and are unable to take input from users for improvements; so need more collaboration from others to improve results.

Q3: How can we ensure that our environmental data management practices remain sustainable—both environmentally and ethically—in a rapidly shifting global context?

- re. getting commercial users to pay for their access to help pay for the services for others.. would be nice but it may not be possible to charge due to policy ... and also showing such usage may actually support 'impact' which is one of the metrics that can help secure core funding... so do we 'cut off our nose to spite our face'?
- technology evolutions lead to decisions that leave others behind. There is little scope for core funding to support more democratic access (e.g. allow access to local data processing or to support dispatching of data any more). Is there a space for other actors to come in and facilitate the access that is needed for true democratised access?
- Responses (post sticky notes anywhere):
- there are data users now (bot driven) that are much hungrier on resources than before... this presents new challenges in terms of sustainable and equitable service provision
- its also about driving more environmentally aware data analysis approaches - "code lean to code green"
- there's a tension in a society/AI/M context that wants/demands immediacy (of data access) vs the needs for greener (colder) storage mechanisms
- There needs to be a structured process to determine what (non-measurement) datasets can be aged off of repositories over time -simulation outputs, new versions of reprocessed measurement datasets. In many cases its much cheaper to preserve the codes used to generate data (e.g. simulation outputs or statistically reprocessed datasets) vs preserving the data indefinitely. See: https://modeldatarcn.github.io/
- We need to make sure that data management skills are being taught to all students at university, not just into HDR programs. I'm studying (part time) a Master of GIS and not once has anyone mentioned deleting data that you no longer need. As an example, you complete one assignment that's used satellite data and then you submit and you're on to the next one. This is a similar pattern to research. We need to break that cycle.
- Be judicious about what data to preserve. Do all "edge use case" parameters need to be preserved for the long term?
- Leverage compression and lossy compression technologies
- store data to the precision level that is actually meaningful
- Life boat scenario - have plans to store data on removable media that can be stored in a physical location and recovered at a later date.
- Employ emerging technologies to amplify the capabilities of limited data curation staff.

- Have plans in place to triage prioritized datasets across partners. 1. Irreplaceable measurements 2. Statistically processed derivatives of irreplaceable measurements 3. snapshots of the codes and configurations of model generated products.
- Include consideration for data management needs during the project planning or proposal development phase.
- work has looked at this in climate data, for example, which showed that some parameters could be significantly curtailed in storage volumes by preserving data down to the correct level of precision (e.g. 3 dps, not 9!)

These contributions and key elements from the discussion will be synthesised towards a publication, planned to be integrated with the discussions at a Co-located Session on the 17 October 2025 'The Long-term Data Workflow from Creator to Re-user: How Can This Be Managed Best for Future Benefit for Research and Development of Predictive Tools' (see link to slides here).

Our panellists

Stephen Bird (Queensland Cyber Infrastructure Foundation) brings understanding of the options and limitations of cyber infrastructure support for the present and the future,

Wim Hugo, a former member of the EOSC long-term data retention task force, CoreTrustSeal Board and vice-chair of the World Data System Scientific Committee,

Siddeswara Guru, data services lead, the Australian Terrestrial Ecosystem Research Network, provides experience of the data demands imposed on an observatory and repository.

Greg Maurer of the US-LTER can provide practical ways to achieve consistency in data acquisition across a distributed network of research sites,

Gretchen Stahlman (Florida State) brings expertise in data curation, education and legacy data integration.

Shelley Stall (American Geophysical Union), is a pre-eminent promoter of Open Data practices for researchers and an expert in scholarly publishing,

Tavita Su'a (Pacific Regional Environment Program) provides fundamental understanding of the requirements for building and supporting an emerging data network and repository in the South Pacific.

Outputs:

1.	**Session summary:** *This 90-minute panel session on Tuesday 14 October reviewed critical factors affecting the data lifecycle over the long term. These included the inherent expansion of data over time across domains and knowledge contexts, the effect of emerging technologies, the increasing energy and financial cost of handling, curating, and using data, and who takes responsibility for the curation of data for the common good. We shall use environmental data for research as our domain focus.*

	Invited panellists from various stages of the data lifecycle used their experience to answer three questions followed by a short Q & A with the audience. 1. *How do we balance the promise of emerging technologies with the practical risks of data loss and preservation challenges across the long-term data lifecycle?* 2. *What does true democratization of environmental data look like—and who might be left out?* 3. *How can we ensure that our environmental data management practices remain sustainable—both environmentally and ethically—in a rapidly shifting global context?* *Together with contributions from the audience, there was a wealth of material produced which remains to be synthesised.*
2.	**Key outcomes/takeaways:** *The session helped formulate key themes around the questions posed and participants (panellists and attendees) have been invited to contribute to a paper aimed to describe a clear understanding of current status and provide guidance for future work, such as:* 1. *The Identification of possible roles of new technologies* 2. *The quantification of the cost of high quality, persistent and abundant data* 3. *The identification of risks to data when supported by big corporations, and mechanisms to reduce risk.*

Promotional flyer: