UK Centre for Ecology & Hydrology
National Capability for UK Challenges
#CC01 - 6th May 2025
1.30pm – 3.30pm
90 participants

Community Conversation

Co-Designing Digital Infrastructure for Environmental Science

Report on #CC01: Enhancing Discoverability & Access to Data

Eunice Agyei (Lead author: eunagy@ceh.ac.uk), David Philip Green, Kelly Widdicks, Faiza Samreen, Philip Trembath, Clare Rowland, Matthew Coole, Michael Tso, Marcia Spencer, and Gordon S. Blair. UK Centre for Ecology & Hydrology. 2025.

What did we do?

Our first in a series of online **Community Conversations** focussed on Enhancing Discoverability & Access to Data, as part of the National Capability for UK (NC-UK) Challenges programme. Prof. Gordon Blair and Dr. Kelly Widdicks as the NC-UK Digital and Data Integration Co-Leads opened and chaired the session, which involved three **presentations**, each accompanied by a **Q&A session** and a **panel** discussion. Between the presentation sessions, we conducted a series of **polls**, on 'Discoverability and Access to Environmental Datasets' and 'The use of AI to Discover Environmental Datasets'. Afterwards, we invited participants to complete an opinion **survey**.

Presentations:

Mr. Philip Trembath Senior Semantic Knowledge Specialist (UKCEH)	**Dr. Clare Rowland** Earth Observation Scientist (UKCEH)	**Dr. Matthew Coole** Senior Research Software Engineer (UKCEH)
Introduction to the Environmental Information Data Centre (EIDC) Catalogue, and recent improvements to enhance discoverability and access data.	An overview of the **Land Cover Map dataset** and the new **Spatial Data Explorer (Beta)** tool in the EIDC Catalogue for enhanced access to such spatial data.	An insight into work on **Artificial Intelligence** (AI) for enhanced discovery of data in repositories such as the EIDC.

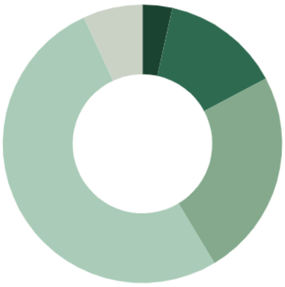
Additional panellists: **Dr. Faiza Samreen** (Software Systems Architect, UKCEH), **Mr. Iain Walmsley** (DevOps Engineering Manager, UKCEH), and **Dr. Christopher Marston** (Ecological Remote Sensing Scientist, UKCEH).

UK Centre for Ecology & Hydrology   NATIONAL CAPABILITY FOR UK CHALLENGES

What did we discuss?

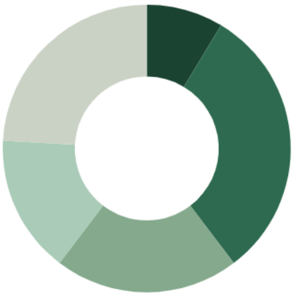
The core aim of the event was to understand more about how users search for, discover, explore, and use data so that we can design solutions to meet their needs within the digital research infrastructure (DRI) landscape. We identified and discussed several **challenges** and **opportunities**:

Discovering data is challenging!



When asked about how easy or difficult it is to access data, many respondents reported difficulties discovering and accessing data for environmental science generally.

■	Very easy	3%
■	Easy	14%
■	Neutral	24%
■	Difficult	52%
■	Very difficult	7%



When asked about whether they use the EIDC Catalogue, some respondents indicated that they have never heard of EIDC Catalogue.

■	I use it regularly	9%
■	I use it occasionally	31%
■	I have used it once/twice	21%
■	I have never used it	15%
■	I have never heard of it	24%

> **“More engagement is needed. User-centred design is key.”**
> **(Participant)**

Discovery across the data lifecycle.

We discussed how access to information about the provenance of data, including the sensors or instruments that collected them, and how data is processed (e.g., cleaning and filtering) helps people understand, trust, and reuse data properly.

Metadata is key.

The conversation highlighted the importance of adding metadata (information about data or datasets) early in the data lifecycle to enable better data management. We also discussed how metadata formats (e.g., Croissant and RO-Crate) support data discoverability, accessibility and reusability by making data machine-readable, semantically rich, and compatible with web standards.

AI offers powerful tools for data discovery when used responsibly.

A key area of focus was AI for data discovery. We shared our prototype search tool leveraging AI techniques – specifically Semantic Search, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) – to support effective and intuitive data discovery and exploration. Semantic search finds the most relevant data by understanding the meaning behind the user's search query, rather than just matching keywords. LLMs then use this retrieved information to generate natural-language answers that are easy to understand. RAG pulls relevant information from specified external knowledge sources to ensure accurate, context-aware responses. It is important that the ethical implications of AI are addressed; our approach and work aim to ensure trust, fairness, and accountability in AI's development and use.

> **"LLM use will support a much-improved user experience."**
> **(Participant)**

Diversify data formats.

The conversation noted how data can be structured in different ways (e.g., JSON vs CSV) and how, by providing data in different formats (especially machine-readable formats), we can help maximise integration, flexibility, and compatibility across diverse systems, and thereby meet the needs of a wider range of users.

Linked data for scalability.

Another important point raised was how federation across datasets can be achieved by using linked data principles (e.g., persistent identifiers), shared metadata standards, protocols, and governance to enable seamless and secure interoperability. Key reference points mentioned included the Australian Research Data Commons (ARDC), for federated data methodologies, the European Open Science Cloud (EOSC) and the broader linked data community.

Interactive maps for accessible spatial data.

Lastly, the discussion emphasised how simple map interfaces can lower the barriers to spatial data with intuitive, visual tools for exploring and filtering information within a geographic context. This means that users do not need Geographic Information System skills or specialised software. For example, the new Spatial Data Explorer (Beta) tool embedded into the EIDC Catalogue will enable users to interact with spatial data via a map.

What would make a difference?

Excellent DRI maximises the value of data. It is important to semantically connect datasets by enriching metadata, building contextual information, and supporting users to navigate and better understand the data. Users can then discover data and insights they may not have found otherwise. The following priorities outline the key enablers to achieve this.

1. **Get the foundations right** by implementing consistent, rich metadata standards that capture context and relationships, using persistent identifiers, and adopting open, interoperable file formats to ensure data is discoverable, connected, and reusable.
2. **Leverage AI driven search to allow data discovery from federated datasets,** with open APIs to transform data discovery from isolated silos to a seamless and integrated experience that enables meaningful discovery of related datasets across platforms and geographic boundaries.
3. **Design around people, not just data,** by co-designing with users, incorporating feedback, and making data accessible through visual tools like maps would ensure that systems meet real-world needs and are used.
4. **Enhance the visibility of platforms like the EIDC Catalogue** by ensuring datasets and metadata are searchable and optimised for search engines. Promote awareness and use of existing data through effective guidance, training, and outreach efforts to maximise reach and impact.

What's next?

- Enhancing EIDC records with new metadata.
- Further experimentation with Croissant and RO-Crate.
- Initial launch and further development, testing and user feedback on the Spatial Data Explorer (Beta) tool as part of the EIDC Catalogue, expanding beyond Land Cover Map for other spatial datasets.
- Further testing and development of the AI enhanced search.
- Integrating the Spatial Data Explorer (Beta) tool and AI enhanced search for a joined-up approach to digital research infrastructure.

Please Sign Up to NC-UK mailing list via NC-UK Website for the latest updates on digital research infrastructure development, as well as for wider updates on the programme and opportunities to engage.

Special thanks to the **organisers** (Dr. Eunice Agyei, Dr. Kelly Widdicks, Dr. David Green, and Ms. Mary Preston) and most of all the **participants** for making this a rich and engaging conversation. Your valuable contribution propels our progress forward.