Improving the discoverability of research methods and workflows

A deliverable report for the AMPLIFY-EDS project

Authors:

Helen Rawsthorne[1] https://orcid.org/0000-0002-6540-8547

Robin Long[1] https://orcid.org/0000-0003-2249-645X

Philip Trembath[1] https://orcid.org/0000-0002-3690-2941

Affiliations:

1 - UK Centre for Ecology and Hydrology

Date:

26 November 2025

.

Contents

Context

We believe this report will be useful because it addresses the growing need for digital research outputs (such as datasets, workflows, and software) to be discoverable and reusable by both humans and machines. By implementing the Digital Objects Ontology (DOO) and standardised metadata practices, we improve transparency, traceability, and interoperability, ensuring that research outputs comply with FAIR principles and deliver long-term value to the scientific community.

About the Environmental Data Service

The Environmental Data Service (EDS) provides a focal point for scientific data and information spanning all environmental science domains: atmosphere and climate, earth observation, polar and cryosphere, marine, terrestrial and freshwater, geoscience, solar and space physics. The EDS is made up of a network of distributed data centres, with domain specific expertise.

Our main goal is to ensure that environmental data are made available, accessible and re-usable for the long-term in order to fully realise their value. We are funded by the Natural Environment Research Council (NERC), part of UK Research and Innovation, to advise researchers on how to prepare data for long term storage and dissemination.

The EDS is a fundamental part of NERC's digital strategy and works with the Digital Solutions Hub, and other partners, to break down disciplinary barriers and facilitate data sharing beyond academic use.

About the project

The AMPLIFY-EDS project was funded by the UKRI DRI Phase 2 call. It ran from March 2024 to Oct 2025.

Within this project, the EDS developed an end-to-end sensor workflow to demonstrate elements of an EDS data commons framework - initiating the first elements in the data commons roadmap. It delivered live data from several research sensor sources through a common workflow and demonstrated their use within shared tools.

AMPLIFY had two primary aims:

- To co-design and develop a live-data prototype service that delivers environmental sensor data and its associated metadata in a standardised, end-to-end workflow.
- To engage with peer sensor data initiatives and stakeholders across the UKRI Digital Research Infrastructure (DRI) and beyond, ensuring alignment, interoperability, and community input.

This work builds upon previous work undertaken via the ENHANCE and BOOST projects.

Introduction

Environmental science research increasingly produces diverse digital outputs such as software, workflows, and models that are foundational to established research outputs (datasets, articles). These newer digital research outputs must be discoverable and reusable by humans and machines to maximise impact.

To address this, we evaluated the suitability of existing metadata standards (CFF, RO-Crate, Bioschemas) for describing methods and workflows and developed the Digital Objects Ontology (DOO), a linked-data-based metadata schema that enables rich descriptions and connections between research outputs.

These initiatives collectively strengthen metadata practices, improve discoverability, and support the NERC community in delivering transparent, reusable, and impactful research outputs. They are the first, vital step in making FAIR methods.

The Digital Objects Ontology

As the environmental science research landscape grows to include an ever-broader range of research outputs, we need to be able to describe and connect them meaningfully and robustly. This is a key element to make them discoverable and reusable by humans and machines. This has benefits which include increasing the impact and reach of research outputs, showcasing existing infrastructure and encouraging multidisciplinary collaboration.

To improve how environmental research outputs are described and connected in online catalogues, we have created a metadata schema grounded in linked data principles in the form of an ontology. The Digital Objects Ontology (DOO, https://nerc-ceh.github.io/digital-objects-ontology)[1] allows us to describe environmental research outputs with structured metadata that goes beyond the standard metadata fields with which we have all become familiar (authors, description, spatial/temporal coverage, licence, keywords, associated publications, etc.).

For any type of environmental research output, the DOO makes it possible to describe and link elements including:

- environmental monitoring facilities/networks that generated data,
- projects that enabled the research,
- funders/grants that funded it,
- people/organisations responsible for delivering the outputs and their roles,
- other research outputs that informed or contributed to the creation of the output.

In practice this means that a metadata record for a dataset could link to the environmental monitoring facilities at which the data were generated, the workflow used to format/quality check/clean it, the grants/awards that funded the research, and the model(s) that used the dataset as input.

This approach also ensures we can give credit to all those involved, such as the field technician, data curator and project manager who played vital roles in producing and publishing the data.

This enhanced metadata offering comes with a multitude of benefits for environmental researchers, the NERC community and research more broadly:

- it increases the FAIRness of research outputs thanks to better traceability and transparency around their provenance
- it improves the visibility of research outputs by making them discoverable via new avenues
- it gives credit to and enhances the profile of all contributors
- it raises awareness of NERC's monitoring facilities and how they are being used
- it encourages collaboration and trans-disciplinarity by showcasing the people and projects who are using facilities and re-using outputs
- it allows researchers to tell a story about their outputs by linking them to tangible elements such as infrastructure, thereby allowing stakeholders to explore the bigger picture behind a research output

Improving discoverability

Methods[i] developed by the NERC EDS are currently difficult to find. They are stored in a distributed manner, in locations such as the method author's local computer storage, an internal server or on many different source code repositories. As there is no single dedicated place for them to be stored with appropriate and standardised metadata, discoverability is severely limited.

[i] NOTE: We use the term "methods" to encompass a wide range of digital research outputs including, but not limited to, computational workflows, code, software, algorithms, models and computational notebooks.

We decided to concentrate on two key points to improve the discoverability of NERC EDS methods:

The first is to ensure that authors of methods store their methods in a source code repository that is under institutional control rather than under personal accounts. For example, the NERC Environmental Data Service GitHub organisation (https://github.com/NERC-EDS). This ensures better visibility, sustainability, management and institutional support.

The second is to encourage authors of methods to provide a standardised set of metadata to properly describe their method. This makes it easier for the method to be correctly cited and subsequently discovered by others.

These two points go hand-in-hand as the metadata should be stored in the repository alongside the method.

We recommend for these two key points to be implemented for any method that supports a published research paper, data product or report, may reasonably be reused (for replication, validation or adaptation) or may need to be reviewed or audited in the future.

In short, if the method supports a publication or could be reused, it should be in a public repository with proper metadata.

Metadata for describing methods and workflows

The ability to find or discover workflows, the methods used in them, or the data created from them heavily depends upon how machine discoverable they are. This means creating metadata information that is comprehensive and machine-readable.

Metadata files should be stored in the root of the repository that holds the code. They should be machine-readable, which means that it allows them to be automatically exchanged across repositories, platforms and organisations.

This in turn improves their discoverability, as they can be integrated into other catalogues such as those of the NERC Data Centres with minimal effort.

Rather than creating our own metadata criteria from scratch, we have chosen to encourage use of an existing one. We identified a series of metadata criteria specifications which could be used to identify and describe a method/workflow. These were then analysed evaluated against existing systems for recording metadata:

- Citation File Format (CFF)[2] - plain text (yaml) files with human- and machine-readable citation information for software
- Two specifications based on json-LD format that uses schema.org as its base metadata standard:
  - RO-Crate[3] - a standard for describing research objects
  - Bioschemas Computational Workflow[4] - a flagship policy of ELIXIR (the European life-sciences Infrastructure for biological Information).

We found that none of the schemas covered everything which we deemed relevant. CFF is already widely used and supported by the research/academic software/code community. It is very human readable but is limited in its scope and is difficult to extend. RO-Crate is most effective at making metadata machine-readable and is very flexible. However, it is more technically challenging to create and less human-readable.

While recognising that these metadata are important, we also need to ensure we minimise the burden placed on researchers who need to create them. The use of RO-Crate should be encouraged in the longer term. However, due to the technical nature of its implementation and in order to lower barriers to engagement it is recommended that CFF is used as a minimum. We have produced an initial set of guidelines (https://github.com/NERC-CEH/repo-guidance/blob/main/cff-guidance/citation-cff_guidelines.md) that explain how to create, complete and store CFF files in method repositories to ensure that they are implemented in a standardised way across NERC EDS.

Recommendations

Future developments should build on the Digital Objects Ontology to strengthen method discoverability. They should aim to lower effort for contributors, while establishing an environment for long-term interoperability and automation. Our recommendations are:

- Establish a lightweight Steering Group for future development of the Digital Objects Ontology. This should include representation from EDS Data Centres, and method authors.
- Define an ontology governance policy to include handling change requests, versioning, deprecation and alignment with external standards.
- Coordinate with other initiatives both within the NERC EDS (for example, EDS working groups, 'HIGH5 for AI' project) and externally (for example, Elixir UK).
- Develop an “enhanced CFF” profile that includes links to monitoring facilities, funders and grants and related research outputs.
- Establish a complete RO-Crate profile for describing methods/workflows.
- Offer a "bridge" between CFF and RO-Crate: create tool(s) that convert validated CFF to RO-Crate metadata.
- Create a crosswalk between the Digital Objects Ontology, schema.org/RO-Crate so that implementers are able to see how terms line up.
- Recognising that adoption is social as much as technical, provide additional guidance, training materials and examples to encourage engagement.

References

[1] UK Centre for Ecology & Hydrology. Digital Objects Ontology (Beta). Revision: 0.1.0. Retrieved from: https://ceh.ac.uk/digital-objects-ontology/0.1.0/

[2] The Citation File Format Contributors. The Citation File Format (CFF). Retrieved from https://citation-file-format.github.io/

[3] Sefton, P., Ó Carragáin, E., Soiland-Reyes, S., Corcho, O., Garijo, D., Palma, R., Coppens, F., Goble, C., Fernández, J. M., Chard, K., Gomez-Perez, J. M., Crusoe, M. R., Eguinoa, I., Juty, N., Holmes, K., Clark, J. A., Capella-Gutierrez, S., Gray, A. J. G., Owen, S., … Pataki, B. E. (2025). RO-Crate Metadata Specification 1.2.0 (1.2.0). researchobject.org. https://doi.org/10.5281/zenodo.13751027

[4] (2023) ComputationalWorkflow DRAFT Profile. Retrieved from https://bioschemas.org/profiles/ComputationalWorkflow