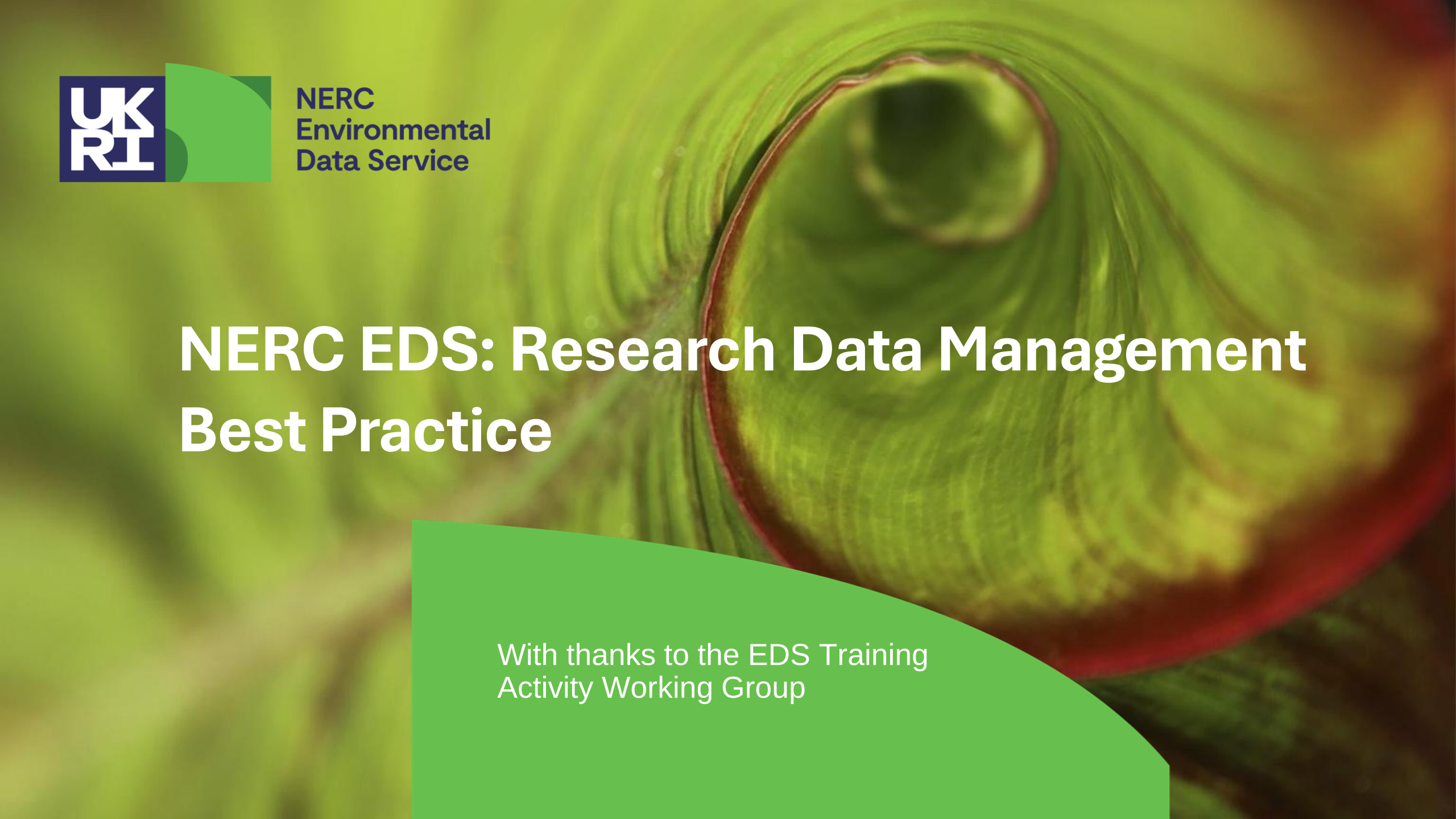


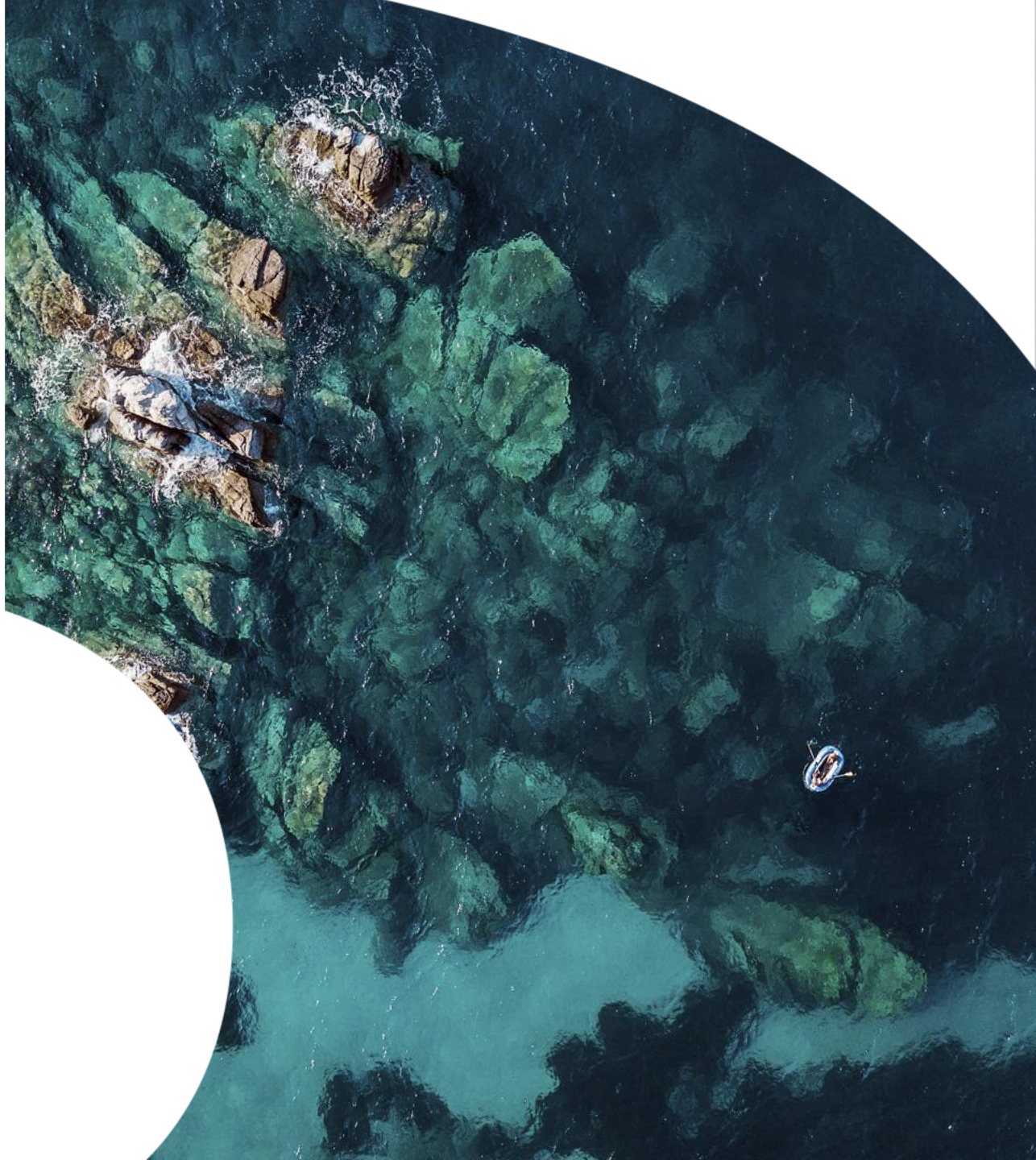
NERC Environmental Data Service

NERC EDS: Research Data Management Best Practice

With thanks to the EDS Training Activity Working Group

Content

- Data Life Cycle
- Metadata and Publishing

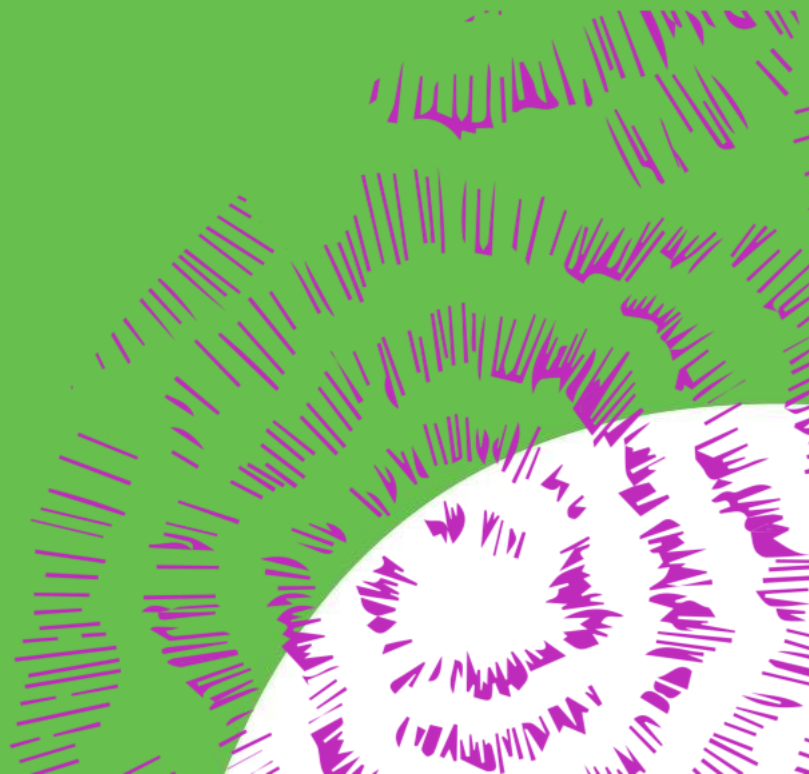


NERC Environmental Data Service

NERC Environmental Data Service

The Data Life Cycle

In this section, you will learn how to organise your data through planning, collection, analysis, publication and beyond

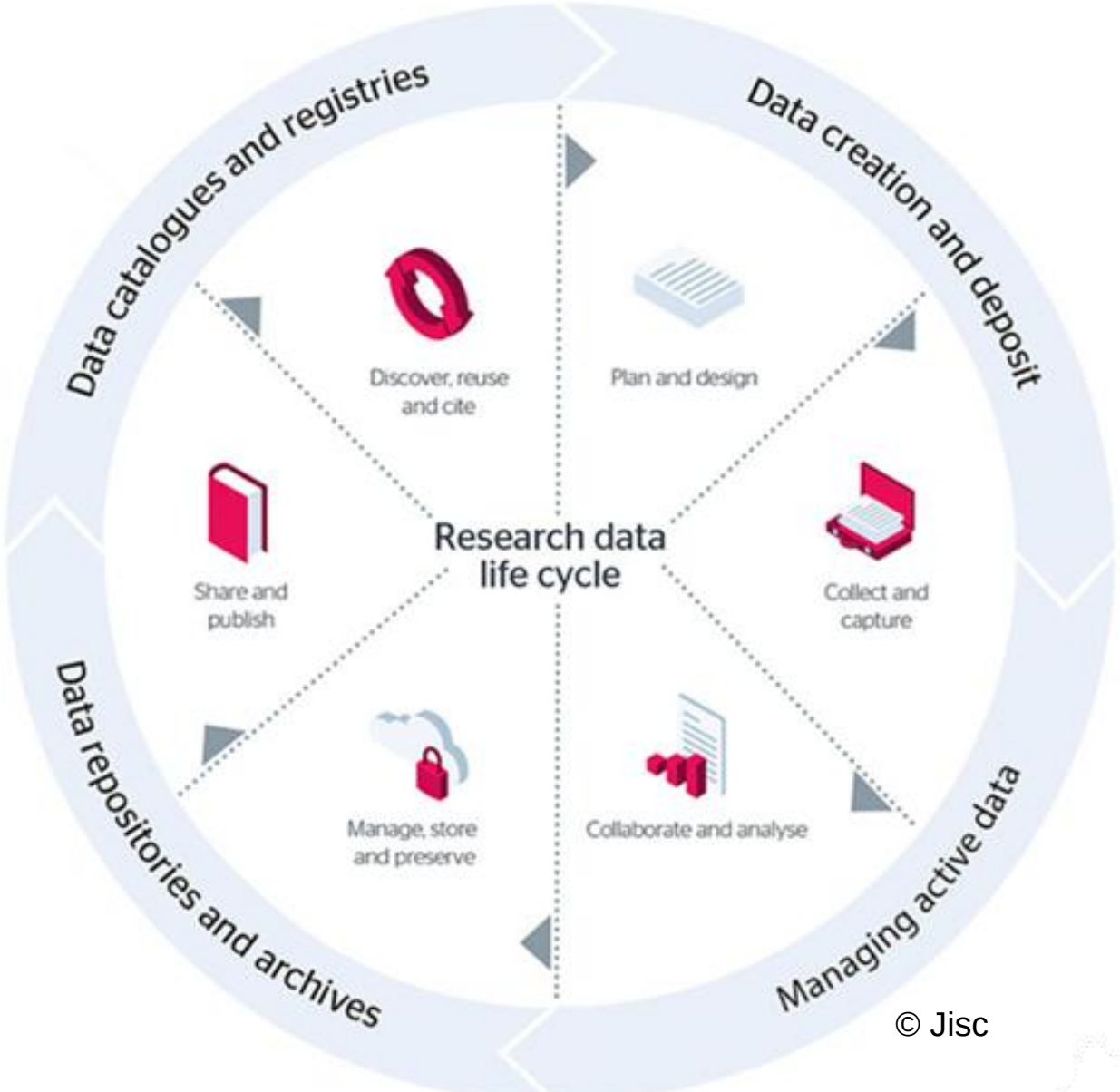


NERC Environmental Data Service

Research Data Management Lifecycle

The data life cycle is the sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion at the end of its useful life.

1) Data Management planning
2) Data collection and capture, collection and analysis
3) Data storage and archiving, sharing and publishing
4) Data cataloguing, discovery and reuse



NERC Environmental Data Service

NERC Environmental Data Service

Depositing Data

Depositing your data with a trusted repository has many benefits including:

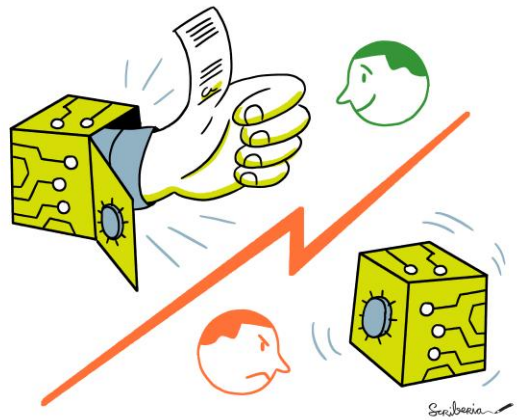
- Ensuring the long-term preservation of your dataset
- Meeting funders requirements
- To gain professional recognition
- To adhere with professional standards and best practice
- To provide access to data users

Check the EDS help page for further information:
https://help.eds.ukri.org/article/5112-controlled-vocabularies

What data should I share and preserve?

You are unlikely to need to preserve all the data you will collect or create. You will therefore need to select data of value and dispose of data of little or no value.

- **What data will be required to validate research findings?**
  - Preserve the raw and final data and the record of processing by which they were transformed from one state to the other.
- **What is the intrinsic value of the data?**
  - Environmental data, for example, are unique to their time and place and have inherent value as part of the historical record. If these are lost they can never be replaced.
- **Are there any restrictions on what data can be shared?**
  - There may restrictions on the data, whether legal, commercial or ethical.
- **Licenses for data:**
  - Creative Commons Licenses
  - Open Government Licence



NERC Environmental Data Service

Checklist for Data Deposit

Metadata

- Will the user be able to understand the data?
- Who? What? When? Where?
- Community standards and vocabularies e.g. NERC vocabulary server

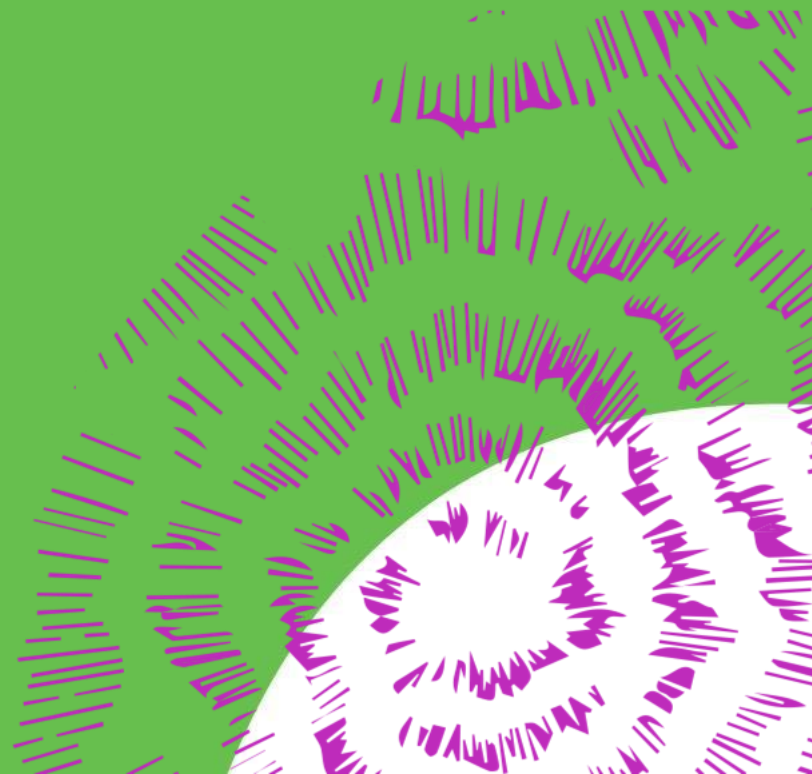
- **Data**
  - Does the user need a specific software to open the data? Is my data in an open format?
  - Did I describe the values with header information (variable name/unit)?
  - Relevant conventions e.g. CF conventions
- **Data transfer agreement**
  - Contact the repository in advance to be ready for my publication. They will be happy to answer any questions
  - Data can be published with embargo and shared with reviewers using an access control system

NERC Environmental Data Service

NERC Environmental Data Service

Metadata

- In this section, you will learn what metadata is and why it is important



NERC Environmental Data Service

Metadata

Metadata: documentation of your data that describes the content, formats, and internal relationships of your data in detail and will enable other researchers to find, use, and properly cite your data.

Information to be given

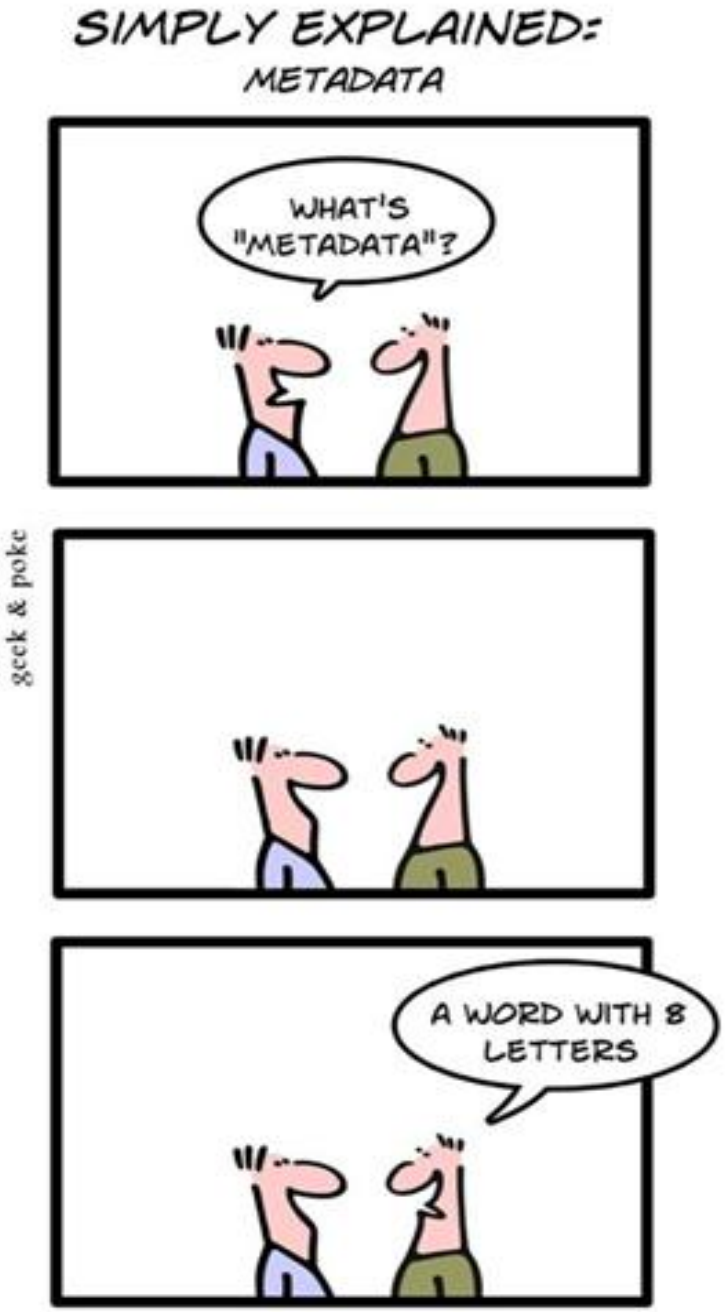
- **WHO** created the data?
- **WHAT** is the content of the dataset?
- **WHERE** have they been acquired?
- **WHEN** have they been created/ acquired?
- **HOW**? Instrument used? Which parameters? Methodology?
- **WHY**? Give the context of the study

- AND ALSO: information about how to use the data (license), how to access it.

NERC Environmental Data Service

Metadata – Data about data

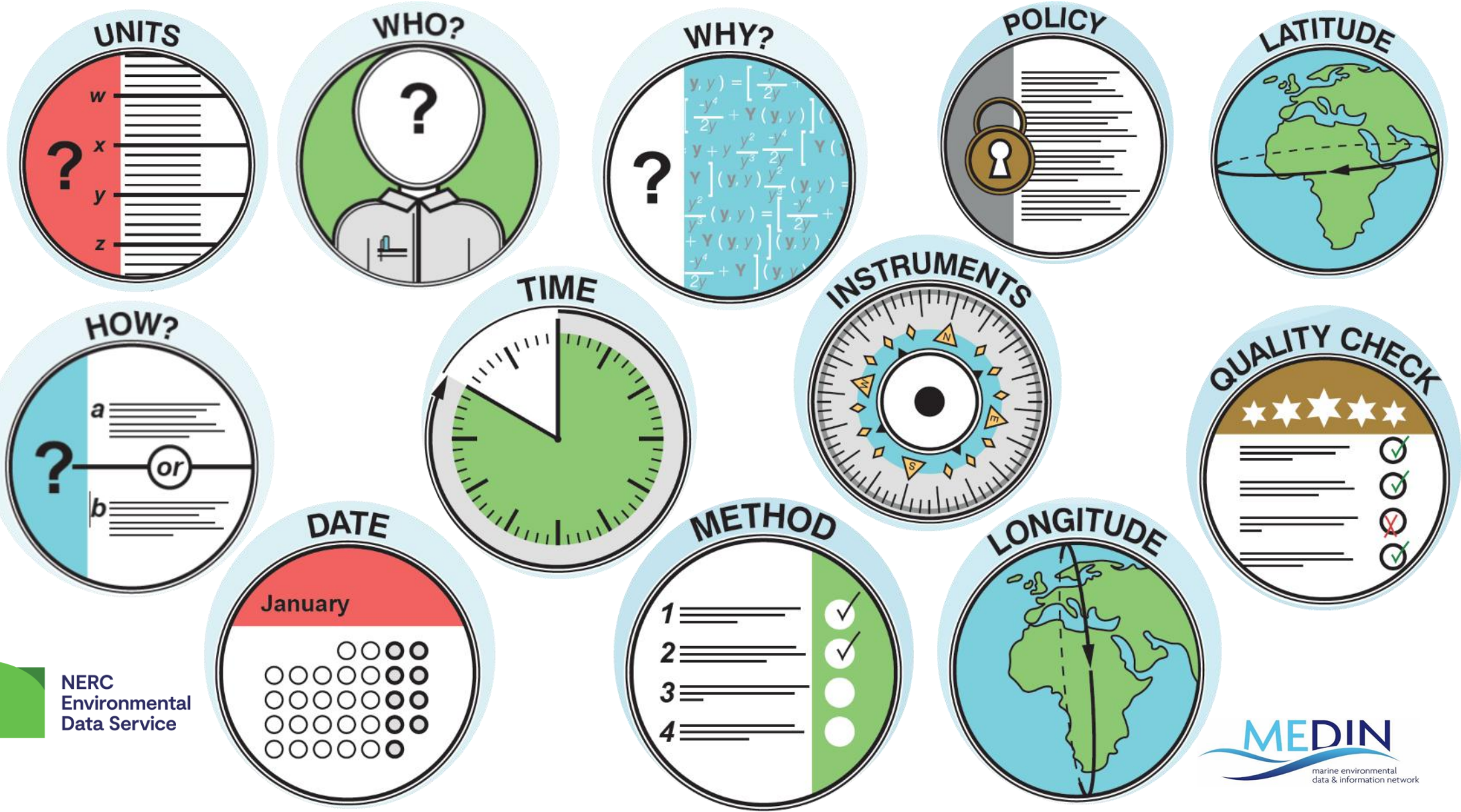
- **Discovery metadata** – enable the data to be found, e.g. experiment name, date, geographical area
- **Browse metadata** – more detailed metadata, e.g., what variables were observed/modelled
- **Usage metadata** – highly detailed e.g. variable names, units, precise coordinates, processing algorithms
- **Citation metadata** – e.g. links to academic papers citing the data, post fact annotations
- **'Extra' metadata** – e.g. detailed metadata about the instrument used



NERC Environmental Data Service

https://help.eds.ukri.org/article/5180-metadata-standards

Metadata – Data about data



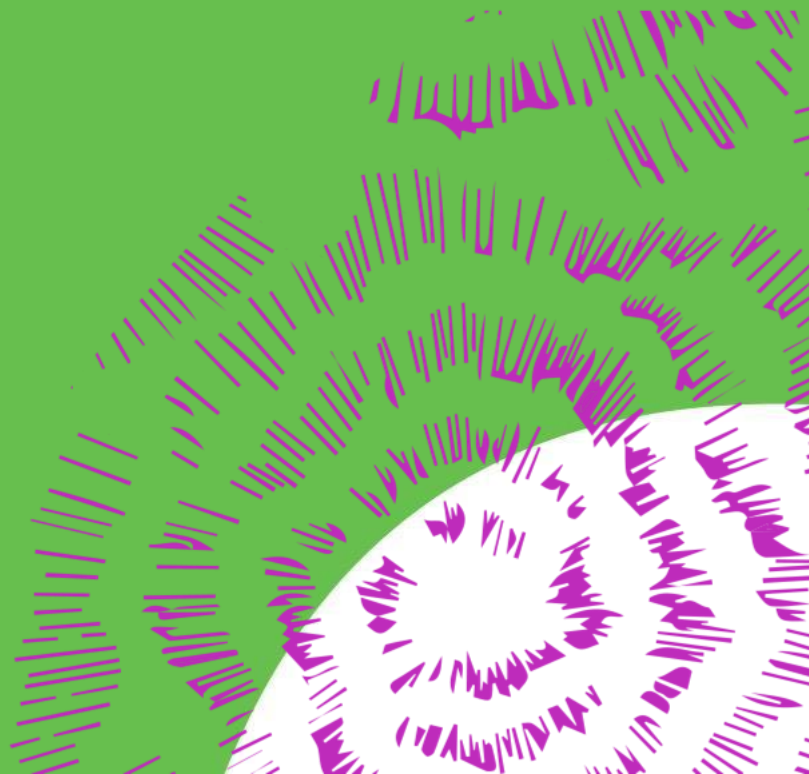
NERC Environmental Data Service

MEDIN marine environmental data & information network

NERC Environmental Data Service

Terminologies

- In this section, you will learn what a controlled vocabulary is and how it can help describe and find your data (and even by machines!)



NERC Environmental Data Service

Terminologies

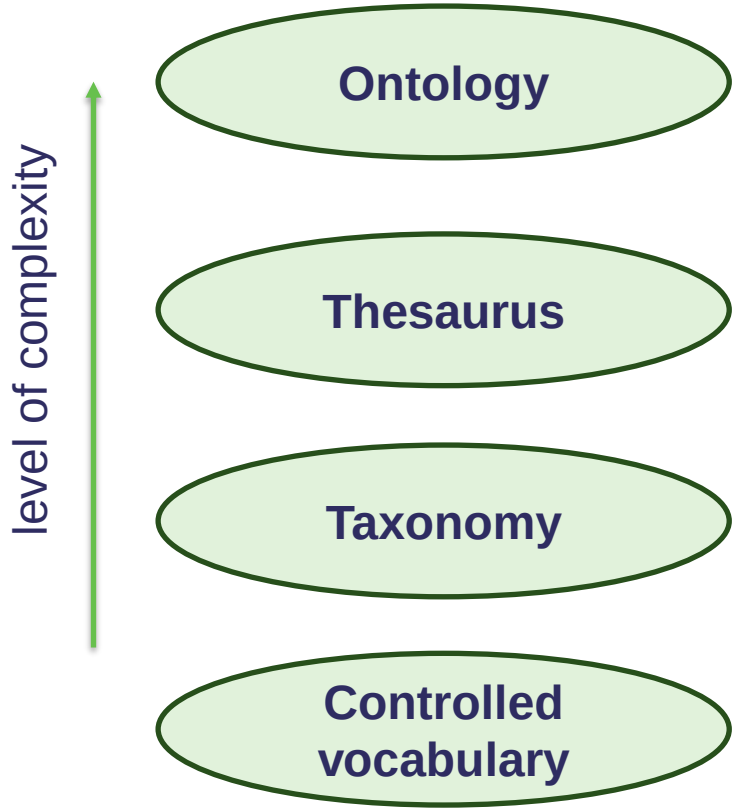
Structured sets of terms designed to enhance the accuracy and consistency of information within a specific domain or scope

- Make implicit more explicit
- Remove ambiguity
- Can enable machine readability

NERC Environmental Data Service

Terminologies

Have **terms/classes/concepts** where each individual element has a unique semantic interpretation and is represented with a unique identifier



a formal version of a thesaurus where relations are described using a formal system such as Description Logic (DL) to mathematically classify individuals of classes and properties, e.g. ENVO: environmental processes and ecosystems.

a controlled vocabulary with a standard structure of terms, where terms have a hierarchical, equivalent or associative, e.g. The Merriam Webster: a thesaurus to find synonyms and antonyms

a controlled vocabulary with a hierarchical structure of terms and a parent-child relationship e.g. World Register of Marine Species: a taxonomy of marine taxa

flat, normalised and restricted list of terms for specific use or context, e.g. NVS: a set of vocabularies from the NERC Vocabulary Server

NERC Environmental Data Service

NERC Environmental Data Service

FAIR Terminologies

Terminologies should follow FAIR principles:

- be discoverable and described with a metadata record
- provide human and machine-readable access to the content
- standardised language to describe terms/classes/concepts
- content has globally unique, persistent and resolvable identifiers
- published in trustworthy semantic repository

Check the EDS help page for further information:
https://help.eds.ukri.org/article/5112-controlled-vocabularies

NERC Environmental Data Service

Publishing data

- In this section, you will learn how to properly publish your data
- You will learn what Persistent Identifiers (PIDs) are and the benefits of use.

NERC Environmental Data Service

RDM lifecycle

- Share and publish data

What to preserve

Licensing

Data checklist

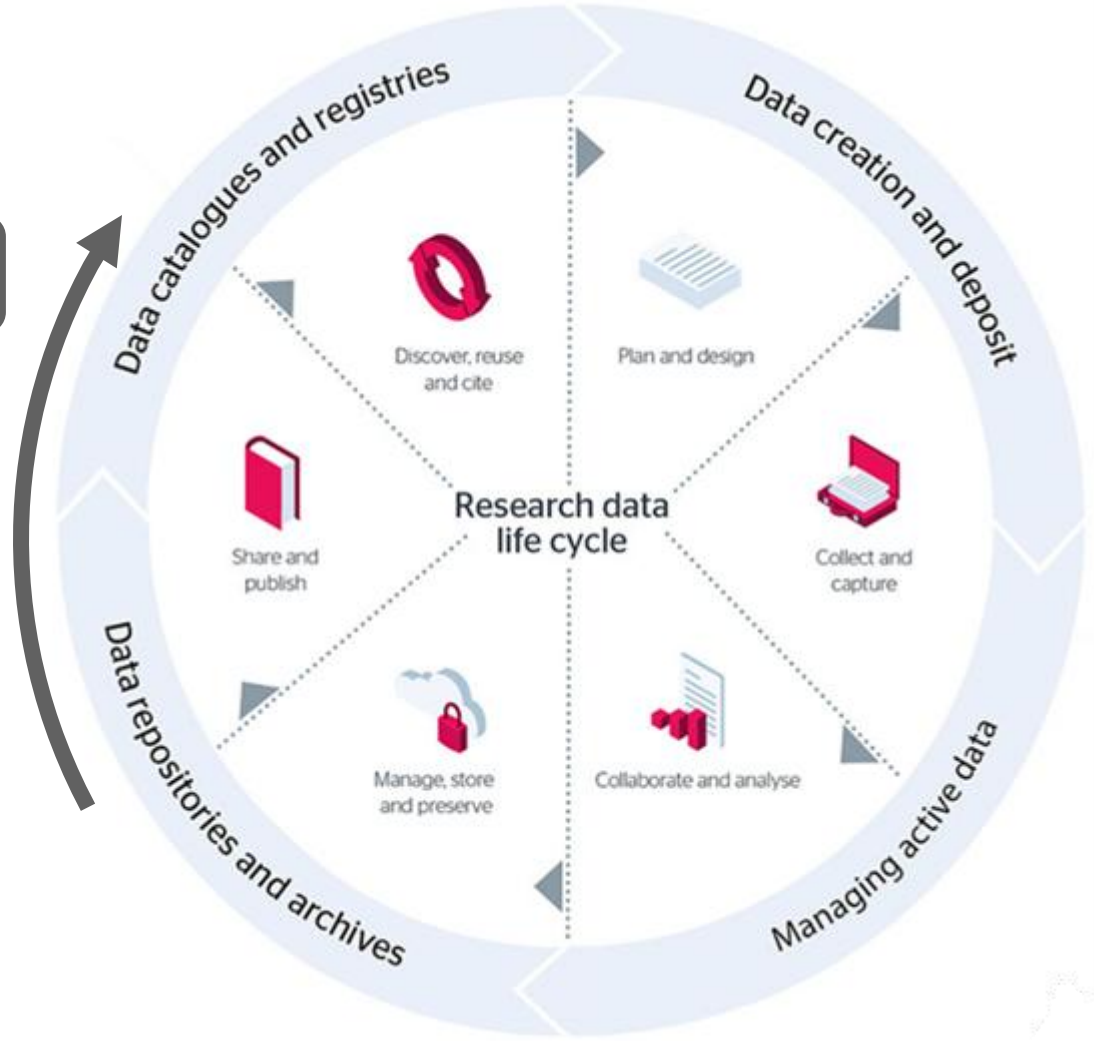
Publishing data papers and code

Persistent identifier

Repository

License

Data Papers



NERC Environmental Data Service

Data sharing: open data licensing

- Data licence is an agreement between a funder and a data user, outlining any limitations on how the data may be used, how the source/creator of the data must be acknowledged and the limits of UKRI-NERC's liability for the data it provides

- UKRI-NERC data licences are usually based on the UK Open Government Licence (OGL) for Public Sector Information

Data reuse rights

- *You can reuse data under copyright but not republish it*
- *Data obtained under a contract has usually more restrictions and contract may supersede copyright*
- *For data which is not clearly licensed you need explicit permission to republish the original data*

NERC Environmental Data Service

OGL

https://help.eds.ukri.org/article/5115-data-licensing-agreements

Data Format

The data format used during analysis might be different from the one used for data publication or data archive

Formats likely to be accessible in the future are:

- Non-proprietary
- Open, with documented standards
- In common usage by the research community
- Using standard character encodings (i.e., ASCII, UTF-8)
- Uncompressed (space permitting)

Examples of preferred format choices:

- Image: JPEG, JPG-2000, PNG, TIFF, geoTIFF
- Tabular data: CSV, (TXT), XML
- Gridded data: NetCDF, HDF5, GRIB
- Audio: AIFF, WAVE

THE GOAL

- Prevent file obsolescence
- For preserved and reusable data

NERC Environmental Data Service

https://help.eds.ukri.org/article/5114-data-formats

Data Papers and data journals

Data paper = a peer reviewed document describing a dataset, published in a peer reviewed journal. It takes effort to prepare, curate and describe data.

Data journals e.g. Nature Scientific Data, Earth System Science Data, Geoscience Data Journal

Case study: KRILLBASE

103 citations… and counting!

Earth Syst. Sci. Data, 9, 193–210, 2017
www.earth-syst-sci-data.net/9/193/2017/
doi:10.5194/essd-9-193-2017

Open Access Earth System Science Data

KRILLBASE: a circumpolar database of Antarctic krill and salp numerical densities, 1926–2016

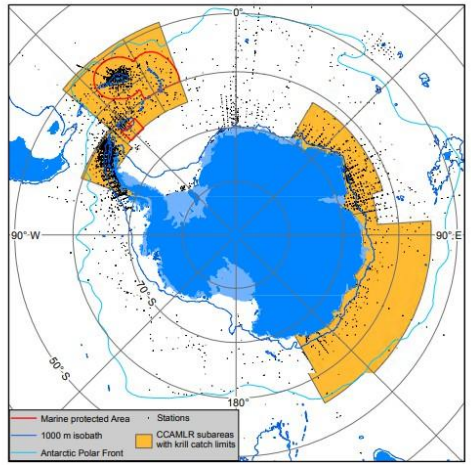


Figure 1. Distribution of sampling stations in KRILLBASE, showing generally elevated sampling effort in and around designated areas of protection and management. These stations may have krill or salp data or both; Fig. S1 in the Supplement provides the distribution of just the krill sampling stations.

NERC Environmental Data Service

Model data

Model output data – *output data of long-term value e.g. data supporting a figure in a publication should be published through a data repository*

Apply guidelines/conventions specific to your model e.g. CMIP6

Ensure you cite the correct version of the model and/or version models you are creating

Code for data should be published via GitHub and Zenodo (see next slide)

https://help.eds.ukri.org/article/5124-model-management

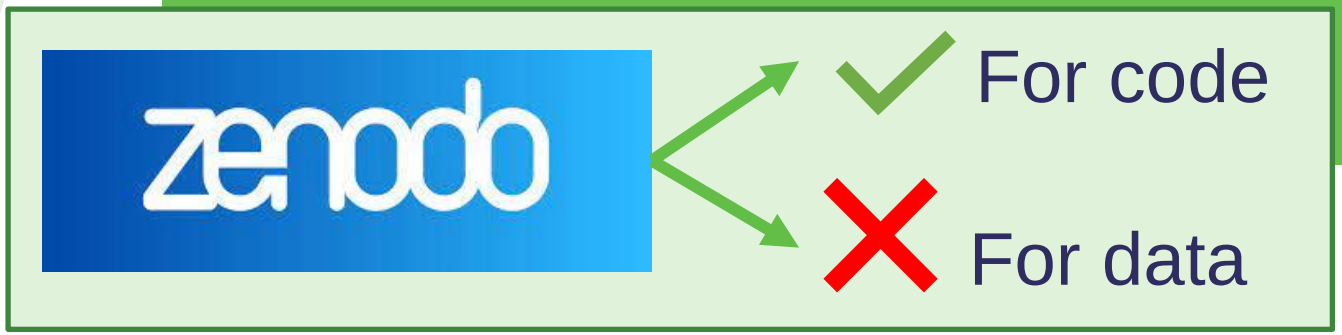
Open code checklist:

- ✓ Is your code well-commented?
- ✓ Does everything execute in order without errors?
- ✓ Does it require human intervention? Does it have a pipeline script which connects all scripts together and executes?
- ✓ Is your code modular?
- ✓ Have you removed local file paths?

NERC Environmental Data Service

Publishing Code via GitHub

- Rapidly evolving code should be open via a public GitHub repository
- Contributors feature allows credit to be given
- Easy to include all the correct documentation
- Push directly to Zenodo to get a DOI for a snapshot of code



Persistent identifiers

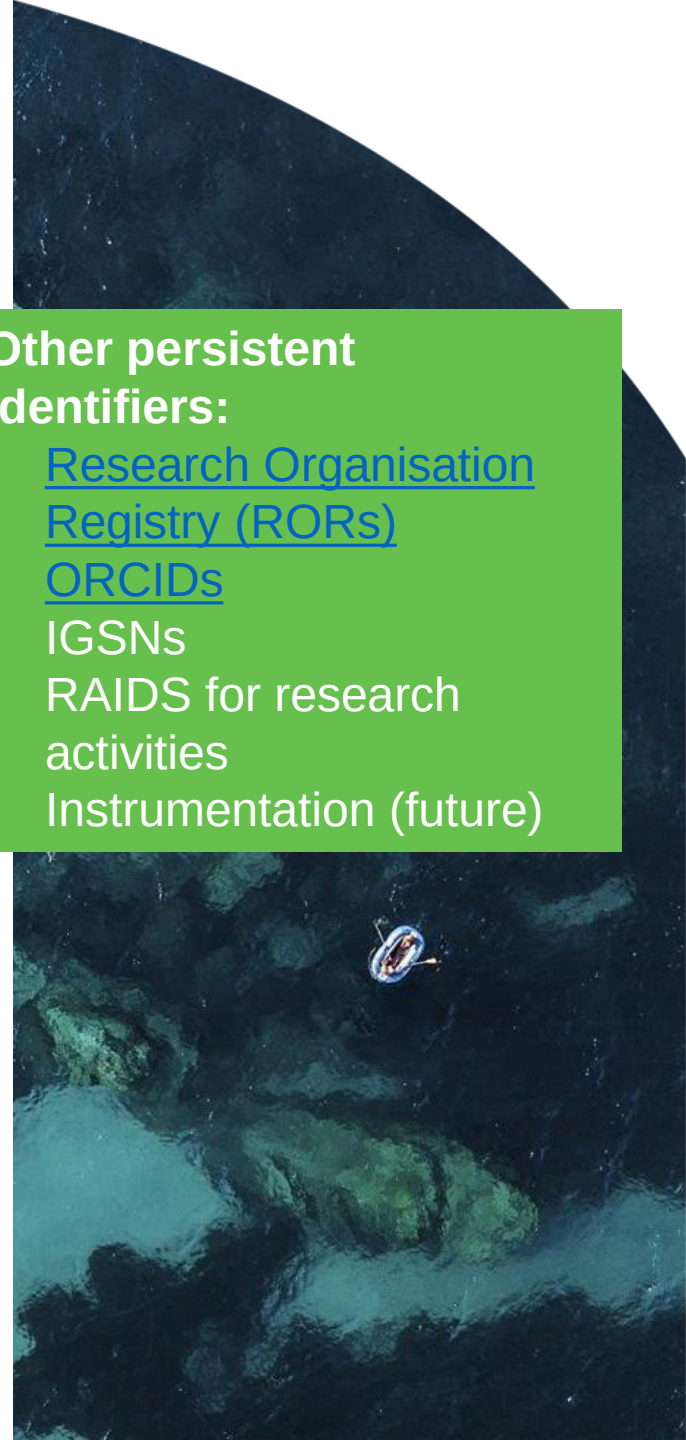
An identifier is a name that you assign to an object, such as a data object. It can be a word, number, letter, symbol, or any combination of those. Ideally, the identifier is unique within your identifier system.

A "persistent identifier" is an identifier that is available and managed over time; it will not change if the item is moved or renamed. This means that an item can be reliably referenced for future access by humans and software. Repositories can provide you with unique, persistent identifiers that you can apply to your data object and manage it over time.

Digital Object Identifier:
Unique and persistent identifier.
DOIs allow better citation of data
and make a direct link from your article
to the dataset and vice versa.
DOIs for datasets are important for linking data reuse and for giving credit for a wider range of people than an academic publication.

Other persistent identifiers:

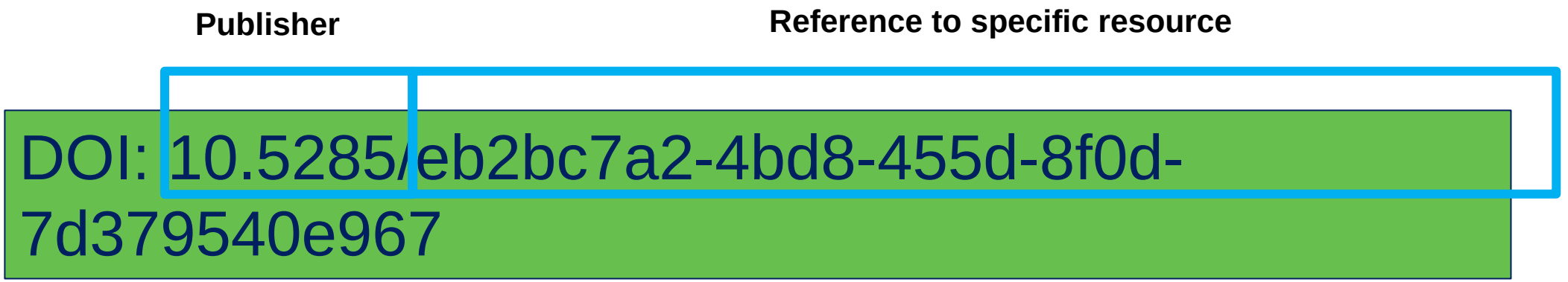
- Research Organisation Registry (RORs)
- ORCIDs
- IGSNs
- RAIDS for research activities
- Instrumentation (future)



NERC Environmental Data Service

Berkeley RDM

DOIs have two components



- two components - ensures long-term accessibility
  - unique identifier comprising a prefix (“10.nnn) indicating the publisher and a suffix identifying the individual article or book
  - service to locate the individual resource

NERC Environmental Data Service

Persistent identifier: Open Researcher and Contributor ID - ORCID

- Persistent digital identifier that distinguishes you from every other researcher
- Alphanumeric code e.g https://orcid.org/0000-0002-1825-0097
- Supports automated linkages between you and your professional activities
- Researcher driven – you need to register yourself
- Recommended minimum information
  - Name
  - Affiliation

ORCID
Connecting Research and Researchers

UKRI NERC Environmental Data Service

Register at: http://www.orcid.org

The Power of PIDs

Published Dataset in a repository

doi

ORCID

Authors and data creators

PIDs allow links to be made between Datasets, Research, and authors

IGSN

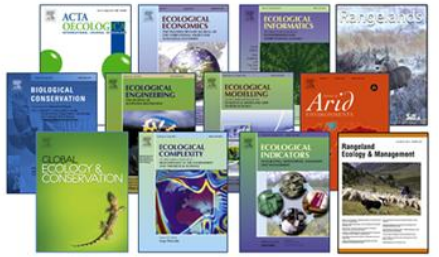
Physical samples

ROR

Research organisations

doi

Research Article or Data Paper



UKRI NERC Environmental Data Service

Data Licences, data citations

- Open vs FAIR – not all data is open. Data should be as open as possible, as closed as necessary (ARDC, 2023).
- - Ensures data has appropriate licences, and any confidentiality processes are followed.

NERC Environmental Data Service

Example of a OGL or bespoke licence

- NERC EDS Team - PublishingData.pdf - All Documents slide 16 – OGL licence

NERC Environmental Data Service