BCO-DMO: Supporting Open Oceanographic Data

Karen Soenen[1], Amber York[1], Shannon Rauch[1], Christina Haskins[1], Nancy Copley[1], Dana Stuart Gerlach[1], Adam Shepherd[1], and Danie Kinkade[1]

[1]Woods Hole Oceanographic Institution

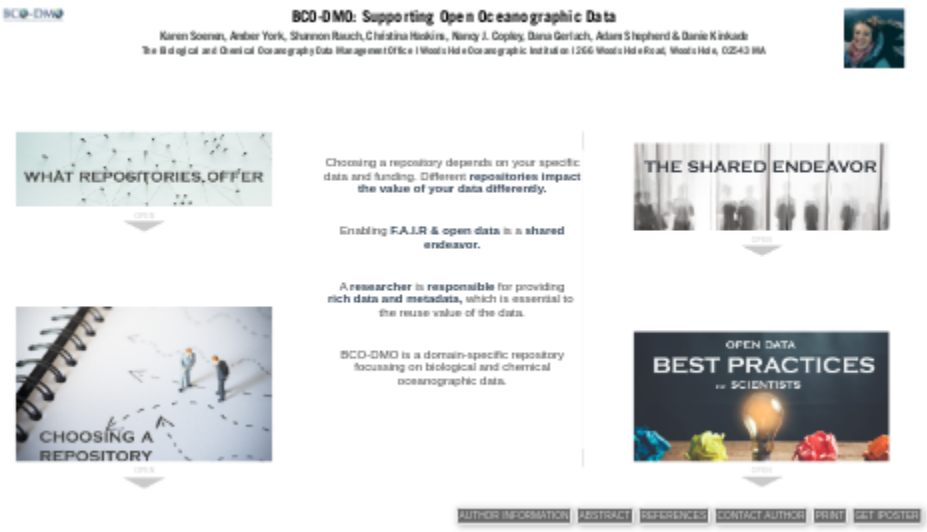
November 24, 2022

Abstract

Open data, data that anyone can access, use, and share have found its place in the current research landscape. Yet, there are many waters to navigate for open access data. Repositories such as BCO-DMO are here to make that journey easier; juggling data principles and policies, funding requirements, publication specifications, research specifics, archiving and discovery through online search engines. BCO-DMO is a domain-specific repository serving highly heterogeneous biological and chemical oceanographic data. The scale of our data ranges from atmospheric aerosols to the bottom of the seafloor, microbes to megafauna. We assist investigators supported through NSF and private funders throughout the Data Life Cycle. We not only have in-house data managers that extensively curate the data and assist submitters in complying with the F.A.I.R Data Principles, but we also develop resources and architecture to tackle technology hurdles associated with these principles to ease data sharing on the behalf of researchers: Designing and implementing ontologies to increase discovery of the data Developing systems to track provenance Creating and improving standard operating procedures to ensure consistent data processing Providing APIs for improved interoperability Developing an automated submission system to ease data sharing burden To make sure that everything we develop addresses community needs, we continue to seek feedback on our processes (i.e what hurdles does the community have with their data and how can we help to overcome them) through a strategic planning committee and continued communication with submitters. We take part in several broader data community initiatives to improve data interoperability and reuse and demonstrate trustworthiness by obtaining CoreTrustSeal certification. How do you determine an appropriate repository to openly share your data? Is BCO-DMO on your list of choices? Let's hear about it.

AGU - iPosterSessions.com

agu2020fallmeeting-agu.ipostersessions.com/Default.aspx



Data repositories collect, manage and store data sets and they are part of the current research landscape. In an effort to provide high quality data, the F.A.I.R principles were developed.

These principles guide repositories to curate data that is optimally re-usable by applying good data management practices and stewardship. The acronym stands for:

- **Findable**: Data are linked to descriptive persistent metadata
- **Accessible**: Data and metadata are open, free and machine accessible
- **Interoperable**: Data and metadata are standardized and use vocabularies. Data points to relevant metadata
- **Reusable**: Metadata are rich, and employ usage licenses, provenance, and community standards.

The tools and concepts that help with the F.A.I.R process that these repositories often apply and use are:

Persistent identifiers: permanent references and unique labels that provide a reference to an object independent of its location.

- Digital object identifier (DOI) which facilites attribution to authors and contributer
- ORCID: A persistent identifier to a specific person

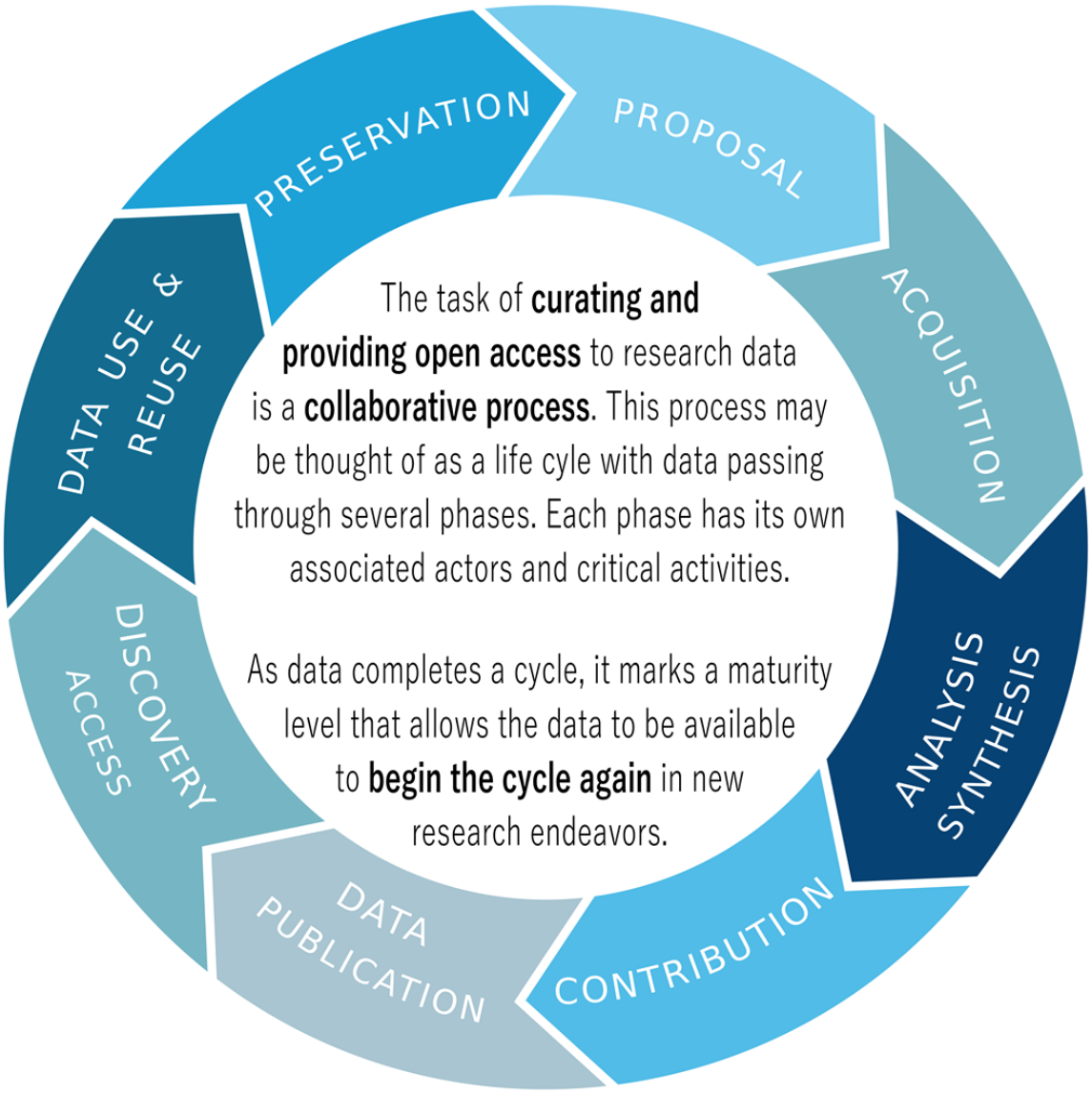
Licensing: A legal arrangement between the creator/depositor of the data set and the data repository, signifying what a user is allowed to do with the data.

Quality control: Checking if the data applies to domain specific requirements and F.A.I.R principles: ensuring datasets' precision and accuracy are well-described, applying consistent formatting to data columns (e.g dates/times)... checking for outliers, etc.

Preservation: Submitting archival copies to an appropriate national data center for long-term preservation.

Online availability: Make data and metadata available online (restricted or public access as appropriate).

Repository managers can act as **data science consultants** for scientists throughout the whole data life cycle and are excellent resources for data management related strategies.


PRESERVATION
PROPOSAL
ACQUISITION
ANALYSIS SYNTHESIS
CONTRIBUTION
DATA PUBLICATION
DISCOVERY ACCESS
DATA USE & REUSE
The task of **curating and providing open access** to research data is a **collaborative process**. This process may be thought of as a life cyle with data passing through several phases. Each phase has its own associated actors and critical activities.
As data completes a cycle, it marks a maturity level that allows the data to be available to **begin the cycle again** in new research endeavors.

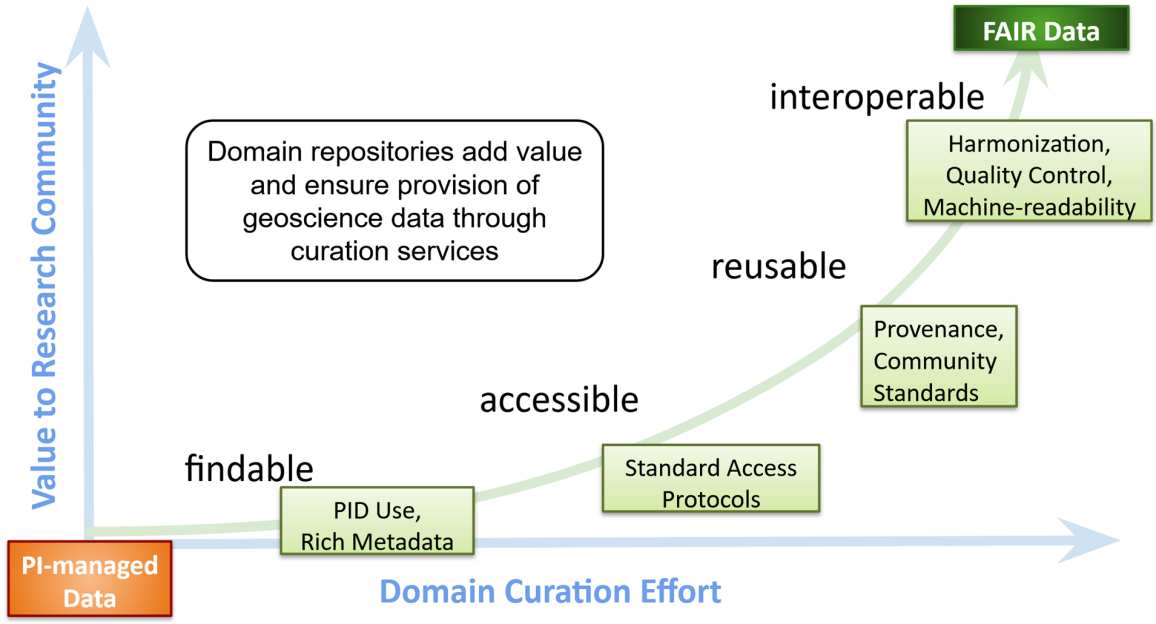
A registry for repositories such as **Re3data.org** (Registry of Research Data Repositories) can help locate suitable repositories for sharing your data.

Yet the ultimate decision to find a repository for your data depends on the following elements:

- **Funder requirements:** Does your organization and/or funding agency have specific repository requirements?
- **Type of data:** Is there a domain repository that your community always uses?
- **Publisher specifications:** Publishers will often request your data to have a DOI, but often do not support the concept of domain specific repositories. Before adding your data to a general repository, look for a domain specific one.

There are 3 types of repositories based on domain specific curration efforts:

1. Domain specific
2. Institutional
3. General



Repository types range from **general repositories**, which curate **heterogeneous types of data**, to Insitutitional repositories who are more familiar with the research at the institution to domain specific repositories.

Domain repositories combine scientific knowledge with information management skills, work closely with their research communities to apply quality controls, create and curate robust discovery- and use-level metadata, and document provenance, thereby increasing data reusability.

In addition, they apply harmonization techniques to data and metadata that increase interoperability across scientific domains. They employ persistent identifiers for metadata and related information, and standardized formats for representing metadata and data; they facilitate the application of appropriate licenses and make connections to qualified external resources for additional interpretation and context for reuse.

Repository Certification

A plethora of repositories are currently available andt the Re3data.org registry can facilitate finding them. Nevertheless, it is often difficult to determine which repositories are capable of fully implementing the FAIR Principles.

This poses problems for data stakeholders seeking a suitable repository for sharing data, and ensuring its access and reuse.

Repository certification such as CoreTrustSeal can demonstrate robust capability and services

The Biological and Chemical Oceanographic Data Management Office (BCO-DMO)

BCO-DMO is a **domain-specific repository** serving research ready data spanning the full range of marine ecosystem related measurements including in-situ and remotely sensed observations, experimental and model results, and synthesis products.

We strive to continuously find solutions to increase data discovery, access and interoperability by improving our technical capabilities, our curation process and actively taking part in the open data landscape discussions and challenges.

We work closely with investigators to serve data and information online from research projects funded by the U.S. National Science Foundation (NSF) Biological and Chemical Oceanography Programs, and the Office of Polar Programs Antarctic Organisms & Ecosystems Program.

Choosing a repository depends on your specific data and funding. Different **repositories impact the value of your data differently.**

Enabling **F.A.I.R & open data** is a **shared endeavor.**

A **researcher** is **responsible** for providing **rich data and metadata,** which is essential to the reuse value of the data.

BCO-DMO is a domain-specific repository focussing on biological and chemical oceanographic data.

Addressing the FAIR sharing principles is a responsibility that is shared between different stakeholders in the data landscape:

- **Scientists**
- **Funders**
- **Federal oganisations**
- **Publishers**
- **Data Community**

Although FAIR data is a mutual goal, there are often different driving factors and potential for discrepancy between requirements and expectations.

All data stakeholders should work together to create a shared publication workflow model and align the goals we all work towards.

Talk by Danie Kinkade on Wednesday December 9 at 4 will focus on this topic during the AGU Fall Meeting 2020.

Publishing data is more than a requirement when authoring an article or a checkbox in your research proposal.

It is an integral part of research and protects your investment in the long term. It is ensuring that your data can be found, accessesed and **re-used in the future without the need for additional input of the data creator**.

No matter the field or domain of a scientist, it is becoming increasingly necessary for researchers to understand and use concepts and tools from the Data Science discipline.

Repos can serve as resources to those scientists, although it is in the best interest of all if continuous effort is made towards data socialization and education about (and usage of) F.A.I.R. data principles.

Described below are important concepts to consider and to apply when acquiring data:

- Data should be both **human AND machine** readable
- Metadata and general documentation
- Controlled vocabularies
- Standards (i.e. time: ISO standard 8601)
- Community best practices
- Non-proprietary formats

Project Organization

Organize your project from the start, use specific naming conventions for both folders and data files and make them meaningful and logic.

Use underscores instead of whitespaces and do not use special characters in the namings to accomodate machine readable structures.

Data Types

What type of data will you be collecting and/or creating? Different data types have different needs, standards, structures.

Different types of data to take into account can be: observations, experimental, models, tabular, gridded, geospatial raw or processed.

Data Formats and Layout

Choose a file format that preferably is usable, open (non-proprietary formats) and readable in the future. i.e. CSV, NetCDF, ...

Data should be both human and machine readable, which for tabular data means specific table formatting is necessary i.e. flat file formatting. Do not use color formatting, but rather create a column specifically for "Notes" that describes the meaning of the color coding.

Geospatial data will need coordinate reference information.

Explicitly use a no data identifier value consistently throughout the dataset.

Metadata - Documentation

Documentation is the most critical part to ensure reusability of your data. Metadata, data about your data, describes the **Who, What, Where, When, and How**. It provides future users with the context to understand and reuse your data.

You should document how your data are collected and processed at every step during the project. Detail where and how it was collected, by whom, which analysis methods were used, and the funding sources.

Metadata should include essential details that another scientist would need to replicate your science.

- Collection dates and methods
- Instrument used (with information about model and settings)
- Were errors encountered, and how were they remedied? What troubleshooting was needed?
- What quality assurance/quality control metrics were used?
- What software (including versions) and what scripts were used?

Standards, Controlled Vocabularies and Community Best Practices

Describe the variables with units and their meaning, where possible use controlled vocabularies.

Dates are a special variable type and the preferred notation is the ISO standard 8601.

Where possible use **community best practices** (i.e. a methodology repeatedly used) to acquire your data and ensure a streamlined dataset in line with others. An example is Ocean Best Practices

Preservation: Think of the preservation of your data, which repository will the data reside in?

Setting up a **Data Management Plan** will help you making decisions regarding the above concepts. The DMP Tool is such tool that can help setting up a solid data plan that can be used throughout the project by all collaborators.