Data Management and Reporting:

BCO-DMO Data Management Services & Best Practices

Shannon Rauch

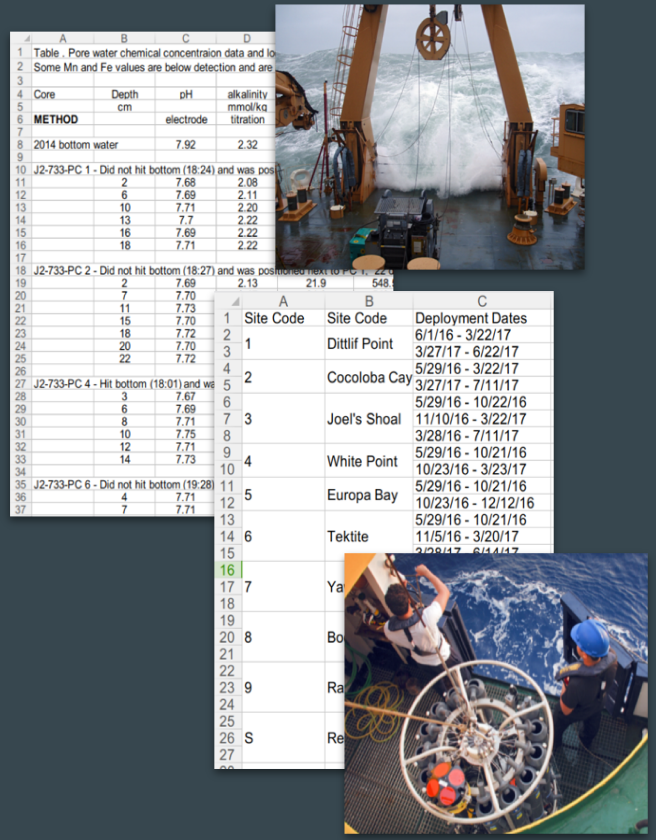
BCO-DMO

NSF
OCE-1435578

Woods Hole Oceanographic Institution

Outline

1. Intro/overview of BCO-DMO
    a. History/mission, who we are
    b. Types of data we handle
    c. Data lifecycle
        i. Proposal:
            1. OCE Data Policy requirements
            2. DMPTool template
        ii. Contribution
            1. How to submit data files & metadata
        iii. Discovery & Access
        iv. Data Publication, Reuse
        v. Preservation
2. Tips for Successful Data Management
    a. Cruise planning & reporting
    b. Event logger (R2R) & sample logs
    c. IGSN & WoRMs identifiers
    d. Data formatting tips/best practices



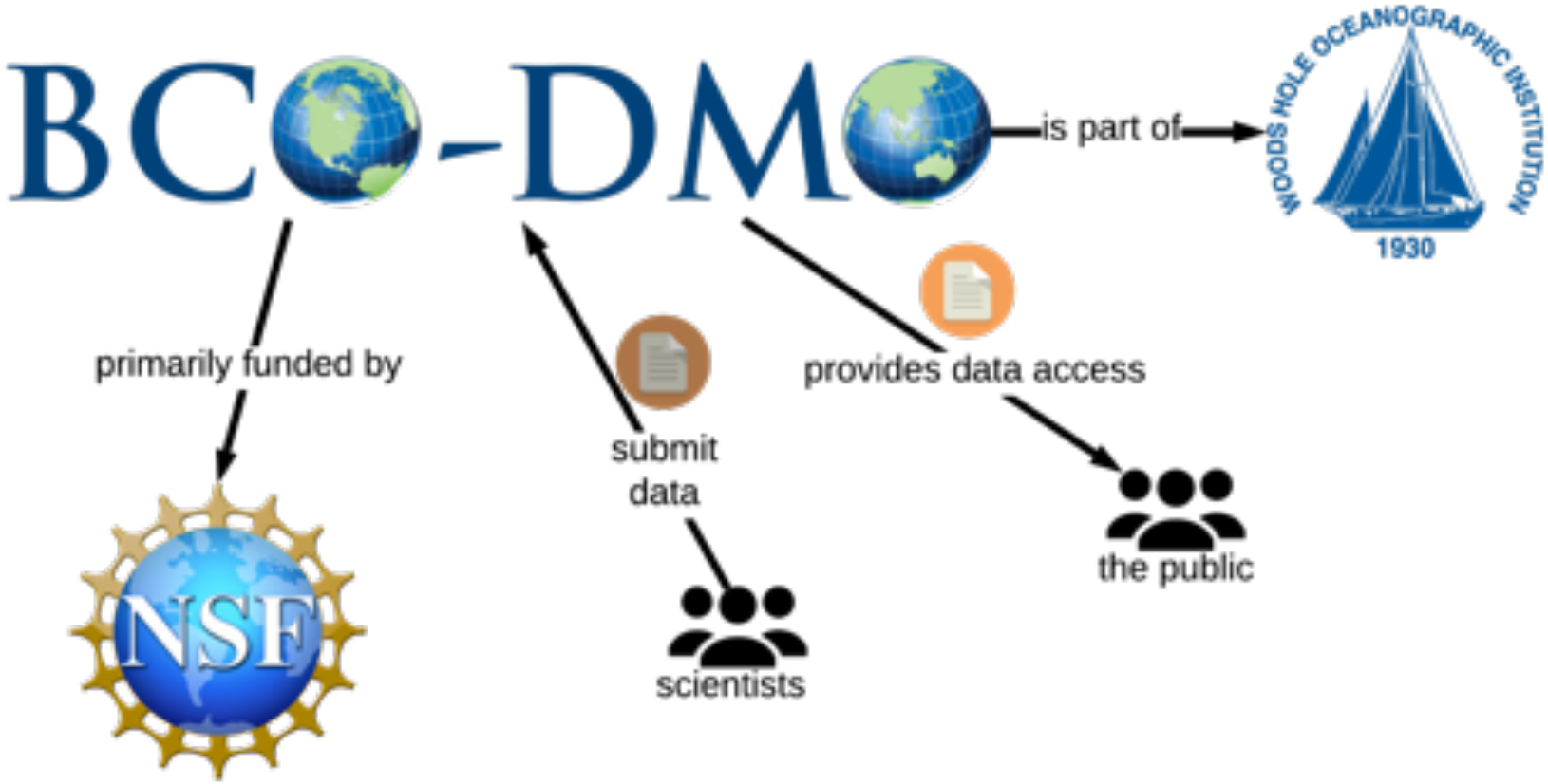
	A	B	C	D
1	Table . Pore water chemical concentraion data and lo			
2	Some Mn and Fe values are below detection and are			
3				
4	Core	Depth	pH	alkalinity
5		cm		mmol/kg
6	METHOD		electrode	titration
7				
8	2014 bottom water		7.92	2.32
9				
10	J2-733-PC 1 - Did not hit bottom (18:24) and was posi			
11		2	7.68	2.08
12		6	7.69	2.11
13		10	7.71	2.20
14		13	7.7	2.22
15		16	7.69	2.22
16		18	7.71	2.22
17				
18	J2-733-PC 2 - Did not hit bottom (18:27) and was posi			
19		2	7.69	2.13
20		7	7.70	
21		11	7.73	
22		15	7.70	
23		18	7.72	
24		20	7.70	
25		22	7.72	
26				
27	J2-733-PC 4 - Hit bottom (18:01) and wa			
28		3	7.67	
29		6	7.69	
30		8	7.71	
31		10	7.75	
32		12	7.71	
33		14	7.73	
34				
35	J2-733-PC 6 - Did not hit bottom (19:28)			
36		4	7.71	
37		7	7.71	

	A	B	C
1	Site Code	Site Code	Deployment Dates
2	1	Dittlif Point	6/1/16 - 3/22/17
3			3/27/17 - 6/22/17
4	2	Cocoloba Cay	5/29/16 - 3/22/17
5			3/27/17 - 7/11/17
6	3	Joel's Shoal	5/29/16 - 10/22/16
7			11/10/16 - 3/22/17
8			3/28/16 - 7/11/17
9	4	White Point	5/29/16 - 10/21/16
10			10/23/16 - 3/23/17
11	5	Europa Bay	5/29/16 - 10/21/16
12			10/23/16 - 12/12/16
13	6	Tektite	5/29/16 - 10/21/16
14			11/5/16 - 3/20/17
15			
16			
17	7	Ya	
18			
19			
20	8	Bo	
21			
22			
23	9	Ra	
24			
25			
26	S	Re	
27			

BCO-DMO: The **B**iological & **C**hemical **O**ceanography **D**ata **M**anagement **O**ffice

- Established in 2006, through the merging of the two data management offices for the U.S. GLOBEC and JGOFS projects.
- BCO-DMO is funded by the U.S. National Science Foundation (NSF) to provide data management services to NSF-funded projects* at no cost to the investigators.
- We now manage data from thousands of projects from researchers (PIs) across the U.S. studying a wide variety of scientific domains.

Who are we?



BCO-DMO: Biological and Chemical Oceanography Data Management Office

Who are we?





Peter Wiebe, Mak Saito, Amber York, Karen Soenen, Nancy Copley



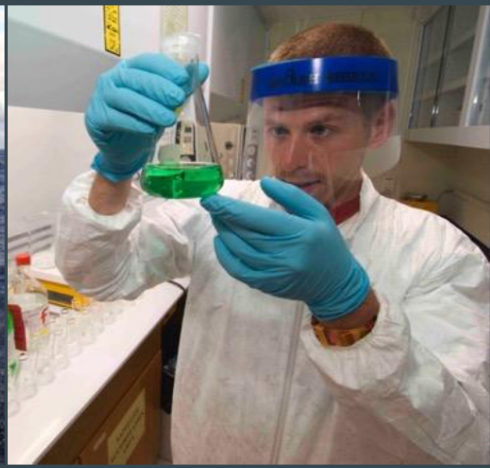


Danie Kinkade, Adam Shepherd, Shannon Rauch, Mathew Biddle, Tina Haskins

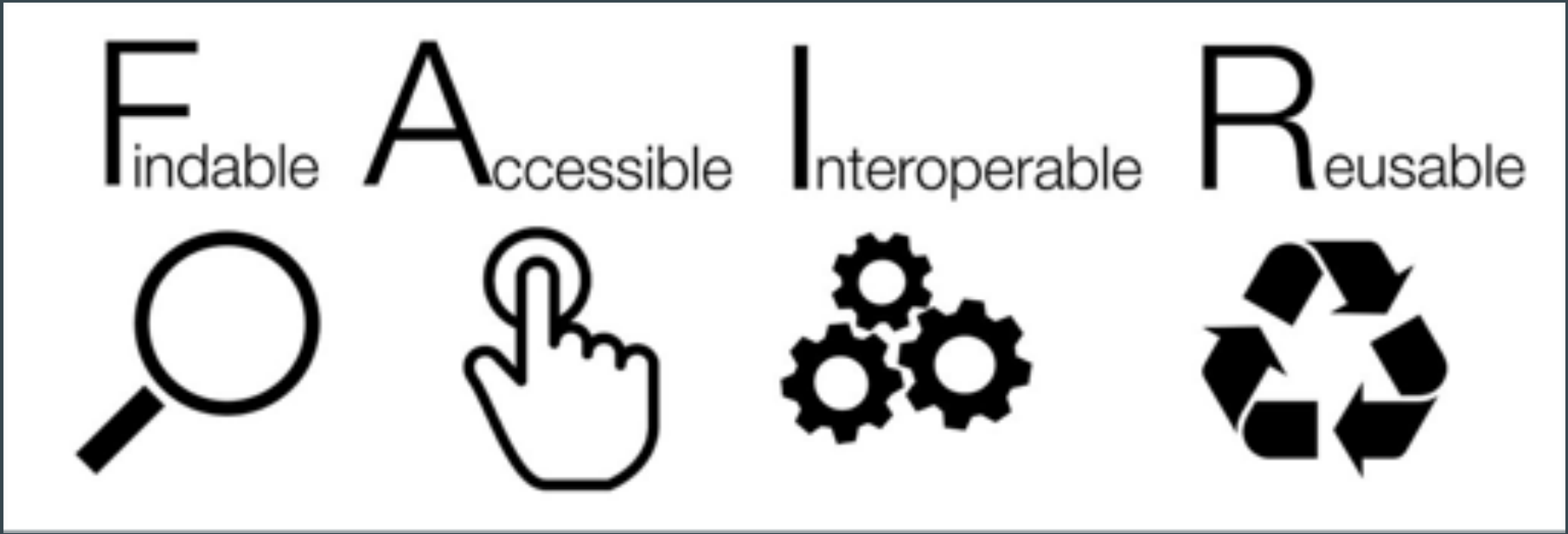
Mission

To work with principal investigators to curate and serve data and information from research projects funded by the National Science Foundation's Biological and Chemical Oceanography Sections and the Division of Polar Programs Antarctic Organisms and Ecosystems program.



Making Data FAIR



Original article describing FAIR data principles: Wilkinson, M.D., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018, https://doi.org/10.1038/sdata.2016.18

Article more specific to FAIR in our domain: Stall, S., et al. (2018), Advancing FAIR data in Earth, space, and environmental science, *Eos*, 99, https://doi.org/10.1029/2018EO109301

At a Glance...

DATASETS = >9000

AWARDS TRACKED = >1900

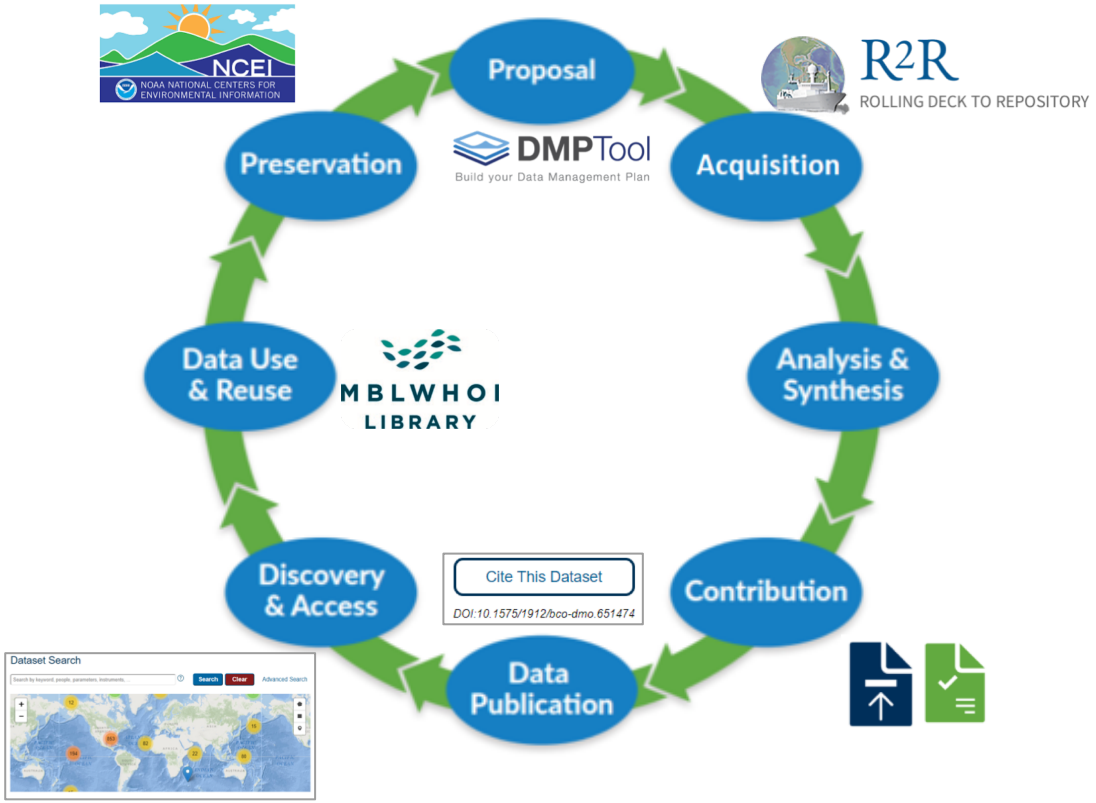
CONTRIBUTORS = 2500

PROJECTS = > 1000

Species Abundances & Biomass
Acoustic Backscatter
Carbonate System Parameters
Chlorophyll & Accessory Pigments
Dissolved & Particulate Nutrients
Ice Cover and Thickness
'Omics-related
Optical Properties
Primary Production
Trace Element Isotopes
Temperature, Salinity, Oxygen

From...the water column, sediment cores, porewater, controlled experiments, etc.

BCO-DMO supports data stewardship throughout the data lifecycle

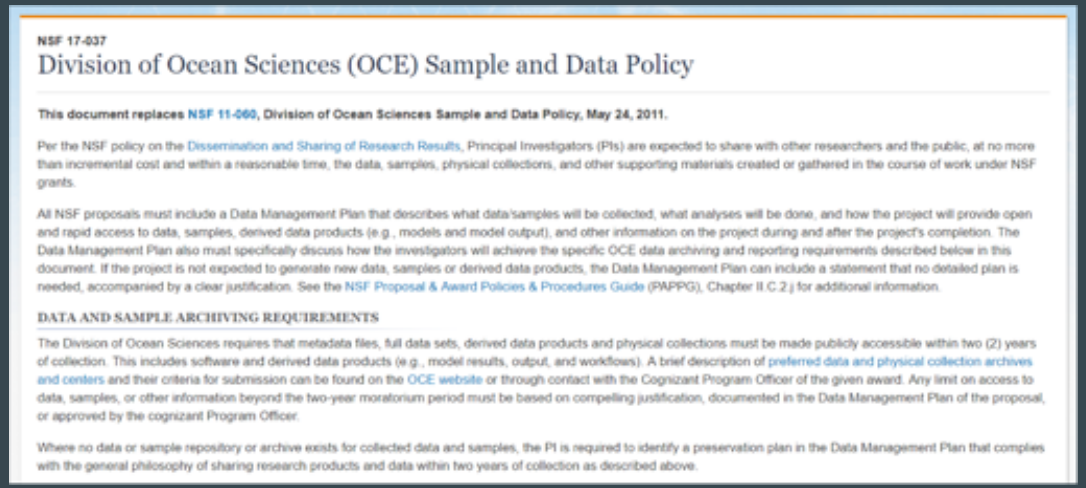


- Provide guidance on data formatting and standards;
- Apply gross QC (e.g lat/lon error checking);
- Capture and record metadata;
- Make data and metadata available online (restricted or public access as appropriate);
- Ensure final archive of data in appropriate national data center;
- Provide Data Management Planning template;
- Assign dataset digital object identifiers (DOIs).

Data Management Planning: NSF Data Policies

Proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan" (often referred to as ‘the two page plan’ or 'DMP'). This document must describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.

Investigators working under awards granted by the NSF Division of Ocean Sciences (OCE) have additional conditions to which they must adhere, as described in the Division of Ocean Sciences Sample and Data Policy.



NSF 17-037

Division of Ocean Sciences (OCE) Sample and Data Policy

This document replaces NSF 11-060, Division of Ocean Sciences Sample and Data Policy, May 24, 2011.

Per the NSF policy on the Dissemination and Sharing of Research Results, Principal Investigators (PIs) are expected to share with other researchers and the public, at no more than incremental cost and within a reasonable time, the data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF grants.

All NSF proposals must include a Data Management Plan that describes what data/samples will be collected, what analyses will be done, and how the project will provide open and rapid access to data, samples, derived data products (e.g., models and model output), and other information on the project during and after the project's completion. The Data Management Plan also must specifically discuss how the investigators will achieve the specific OCE data archiving and reporting requirements described below in this document. If the project is not expected to generate new data, samples or derived data products, the Data Management Plan can include a statement that no detailed plan is needed, accompanied by a clear justification. See the NSF Proposal & Award Policies & Procedures Guide (PAPPG), Chapter II.C.2.j for additional information.

DATA AND SAMPLE ARCHIVING REQUIREMENTS

The Division of Ocean Sciences requires that metadata files, full data sets, derived data products and physical collections must be made publicly accessible within two (2) years of collection. This includes software and derived data products (e.g., model results, output, and workflows). A brief description of preferred data and physical collection archives and centers and their criteria for submission can be found on the OCE website or through contact with the Cognizant Program Officer of the given award. Any limit on access to data, samples, or other information beyond the two-year moratorium period must be based on compelling justification, documented in the Data Management Plan of the proposal, or approved by the cognizant Program Officer.

Where no data or sample repository or archive exists for collected data and samples, the PI is required to identify a preservation plan in the Data Management Plan that complies with the general philosophy of sharing research products and data within two years of collection as described above.

See:
https://www.nsf.gov/pubs/2017/nsf17037/nsf17037.jsp

In short, NSF OCE requirements...

- Full data sets, derived data products (e.g. models and model output), and physical collections must be publicly accessible within 2 years of collection.
- The 2-page DMP should describe:
  - Types of samples, data, software, other materials expected to be produced by the project
  - Standards used to format the data and metadata
  - Policies and provisions for sharing
  - Plans for archiving

BCO-DMO

Biological & Chemical Oceanography Data Management Office

Makes it easier for you to meet the data policy requirements!

Proposal

DMP Template

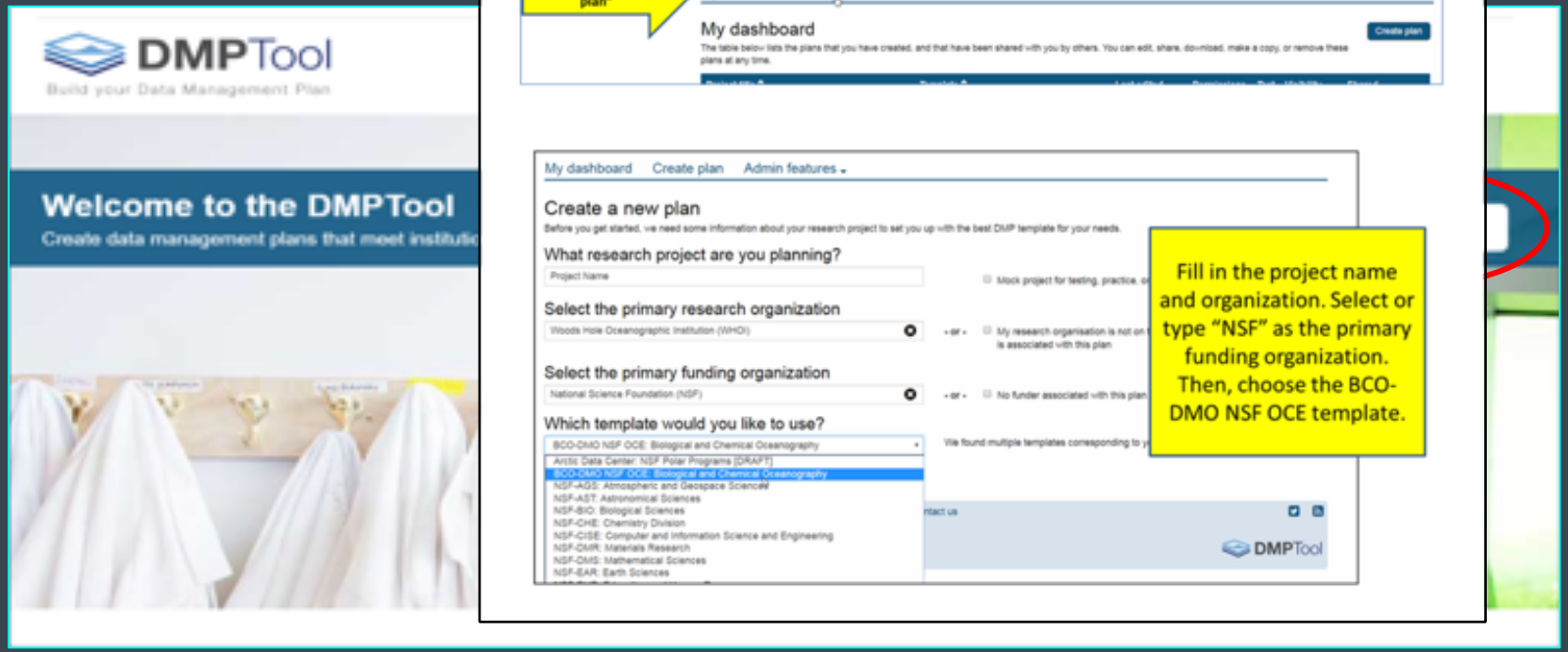
BCO-DMO-created a template designed to meet the NSF OCE Data Policy Requirements

Available through DMPTool.org

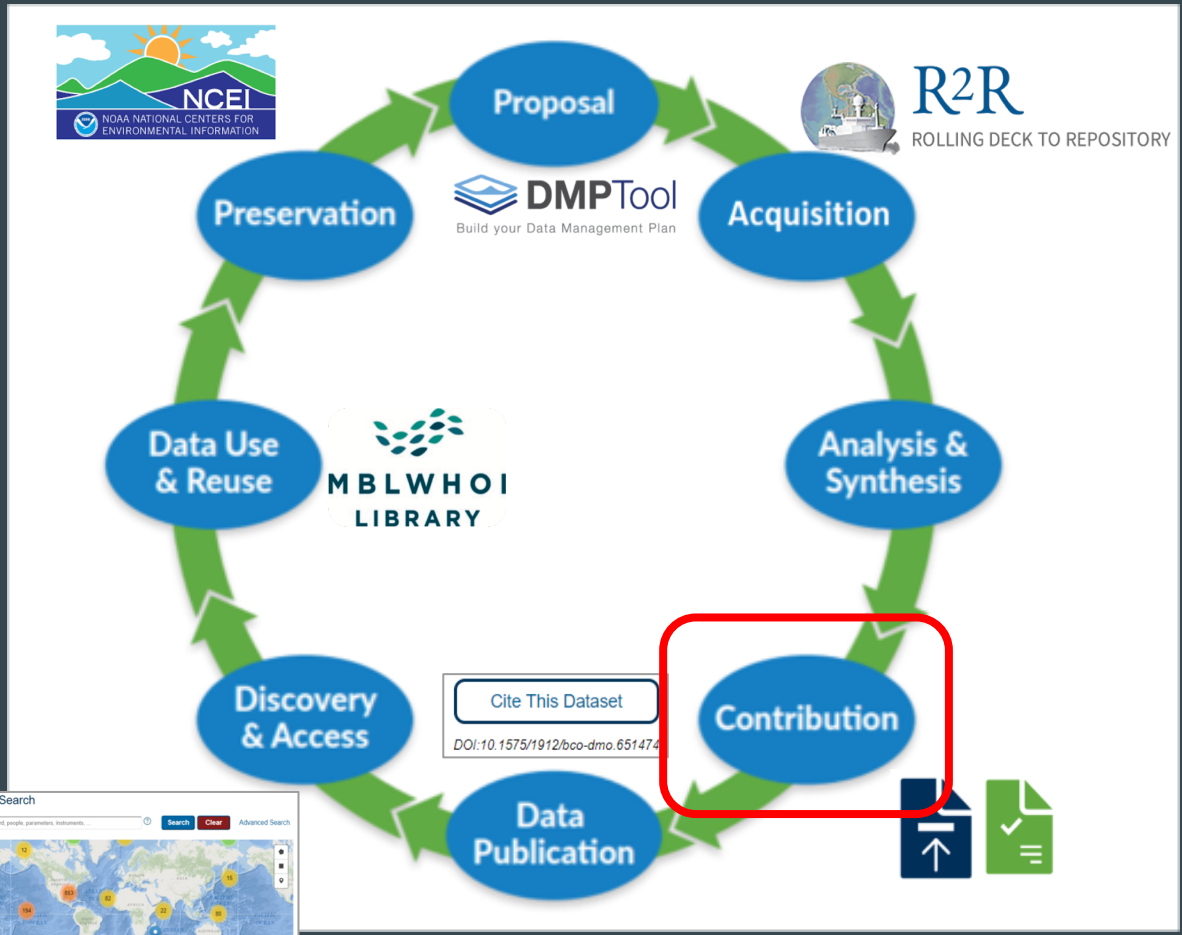
DMPTool is a free, open-source, online application. It provides detailed guidance and walks researchers through the process of generating comprehensive plans.

Plans can be worked on collaboratively, saved, shared, and exported.

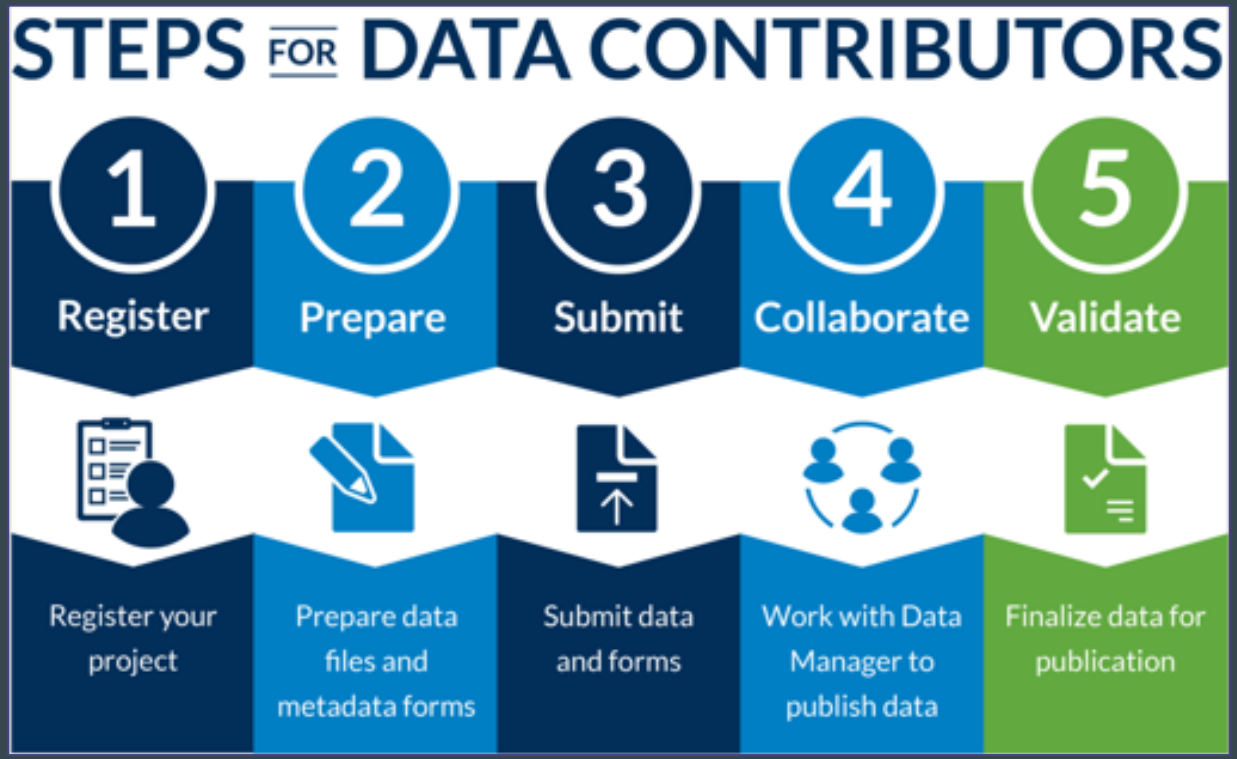
DMPTool



Proposal


NCEI
NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
R2R
ROLLING DECK TO REPOSITORY
Proposal
DMPTool
Build your Data Management Plan
Acquisition
Preservation
Analysis & Synthesis
Data Use & Reuse
MBLWHOI LIBRARY
Discovery & Access
Cite This Dataset
DOI:10.1575/1912/bco-dmo.651474
Contribution
Data Publication
Dataset Search
Search
Clear
Advanced Search

Data Submission Process



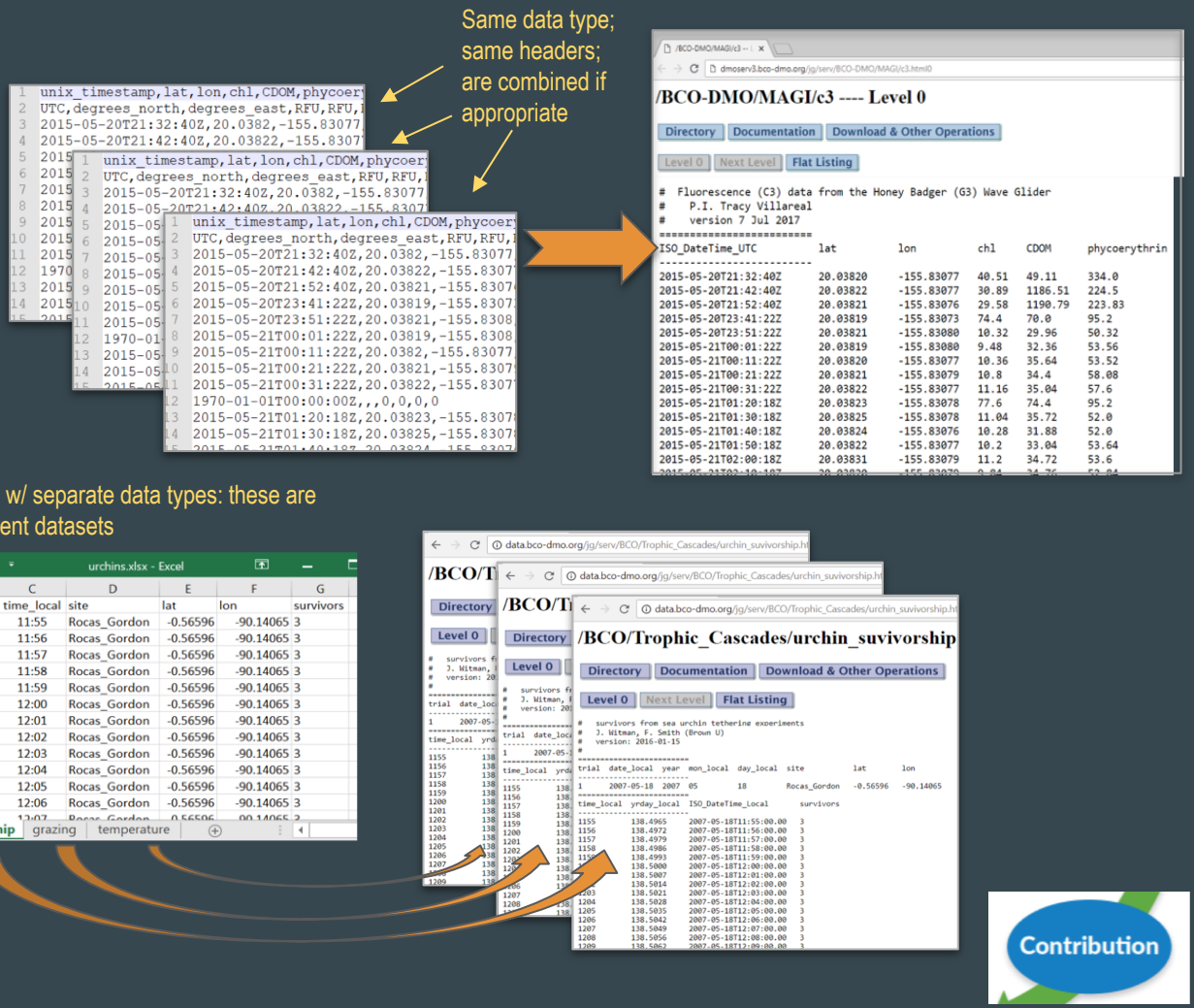
Contribution

Preparing Data

Submit data in the format most appropriate for your community.

BCO-DMO transforms data into a tabular version (csv, tsv).

For videos, images, and other formats not suitable to tsv/csv, we will work with you to arrive at the best data representation possible.



Preparing Metadata

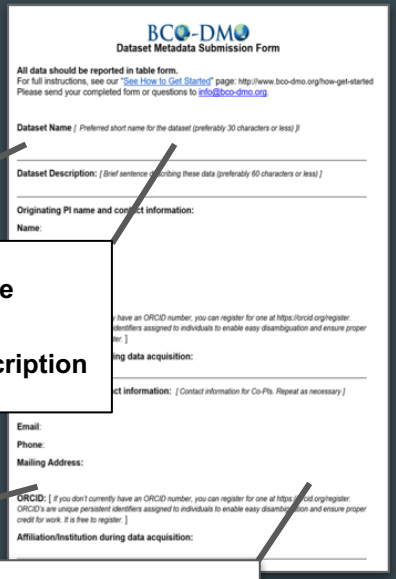
Metadata: describes your data (where and how it was collected, by whom, analysis methods, funding sources, etc.)

The contents of your metadata form are directly used to populate the public Dataset Landing Page.

Allows your data to be understood and re-used by others.

BCO-DMO dataset landing page

Dataset Metadata Form



Dataset Name

Dataset Description

Methodology

Sampling and analytical procedures

Contribution

Metadata Components

- How were the data generated?
  - How were samples collected?
  - How were they processed and analyzed?
    - Methods - include references and citations
    - Instruments - include manufacturer and model; as well as calibration information! (for both lab instruments and shipboard instruments)
    - Be specific; e.g. describe all the sensors on the CTD (not just "CTD")
- Describe the quality control
  - Replicate samples, inter-comparisons, reference materials
- What software or scripts did you use?
  - Are they publicly available?
  - Modeling project? Describe the inputs, parameters, conditions, etc.
- Consider the longevity of your data… they have value beyond your own research!

Submitting Data

Send applicable metadata forms and data files as attachments to info@bco-dmo.org

Talk to us if data are too large for email…

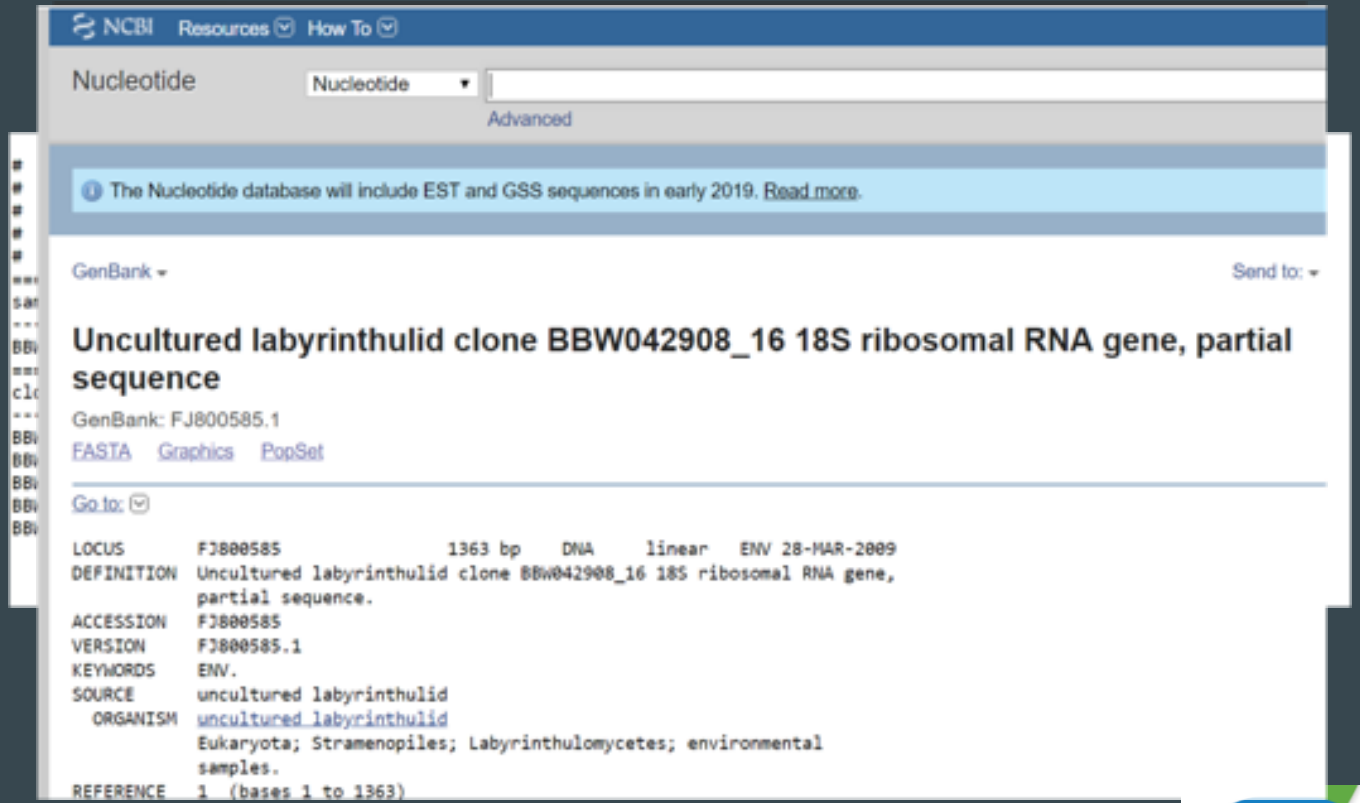
In the works: an auto-submit system. Stay tuned!

Contributing Sequence Data

Sequences themselves are best served by specialized repositories, like NCBI's GenBank.

BCO-DMO serves the related environmental data and links out to NCBI, etc.

This allows all your project data to be discoverable from one place, but served by the domain repositories best suited to do so.

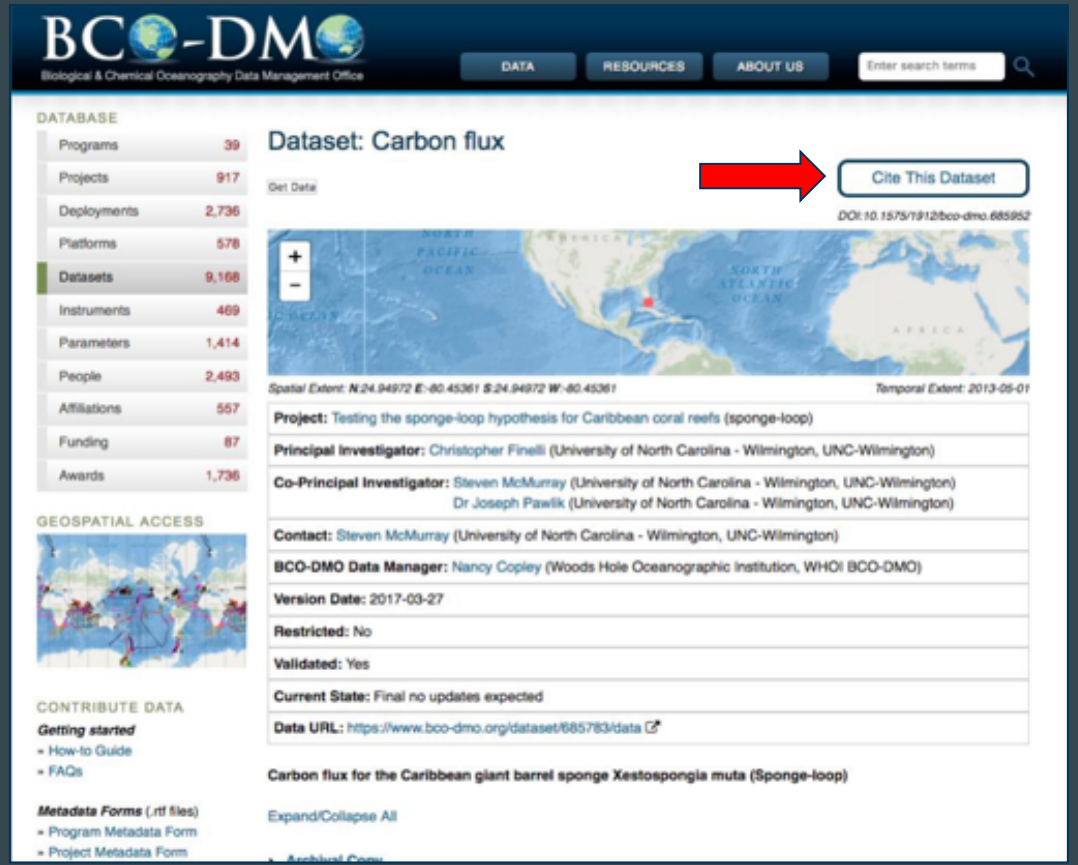


Contribution

Data Publication

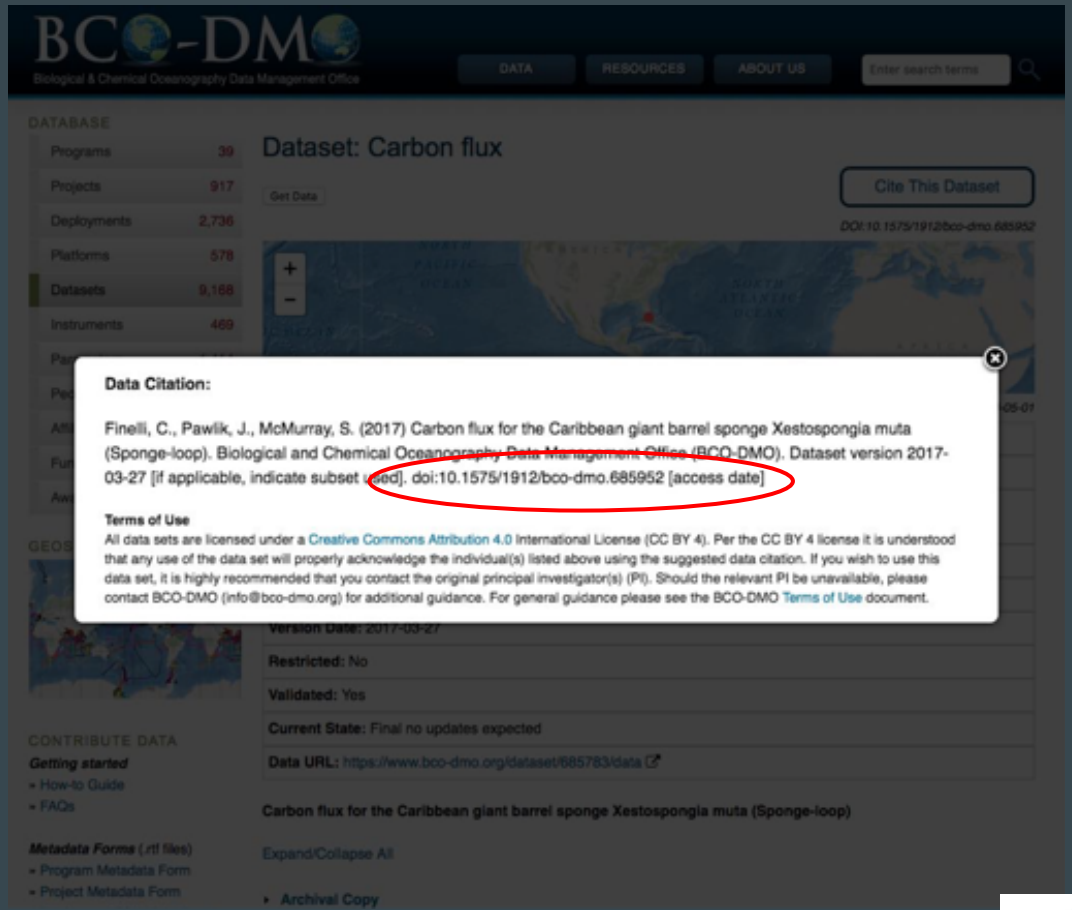
Once data & metadata are validated by the submitter, a Digital Object Identifier (DOI) is assigned to the dataset.

This makes it easier to cite the dataset and discover it (e.g. from a publication to the repository).



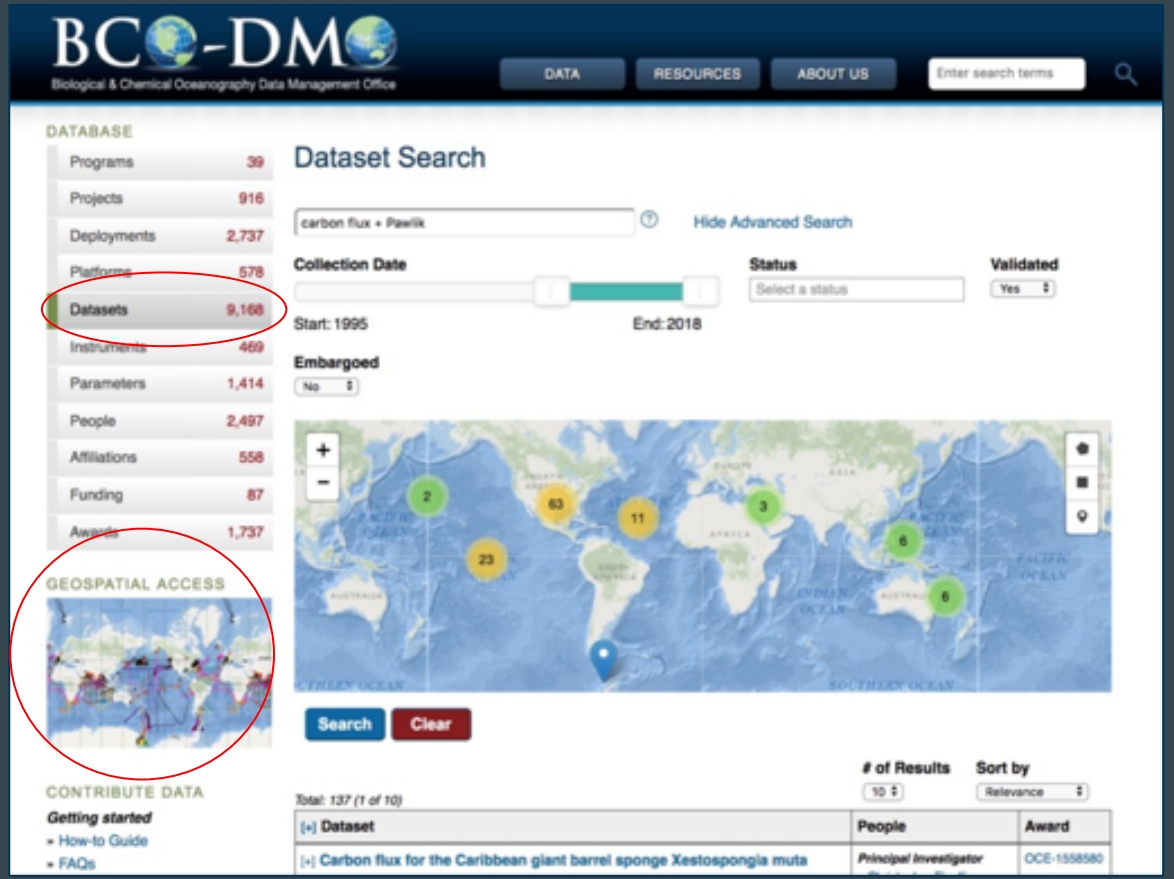
Data Publication

A suggested citation format is provided, supporting data re-use.



Data Discovery & Access

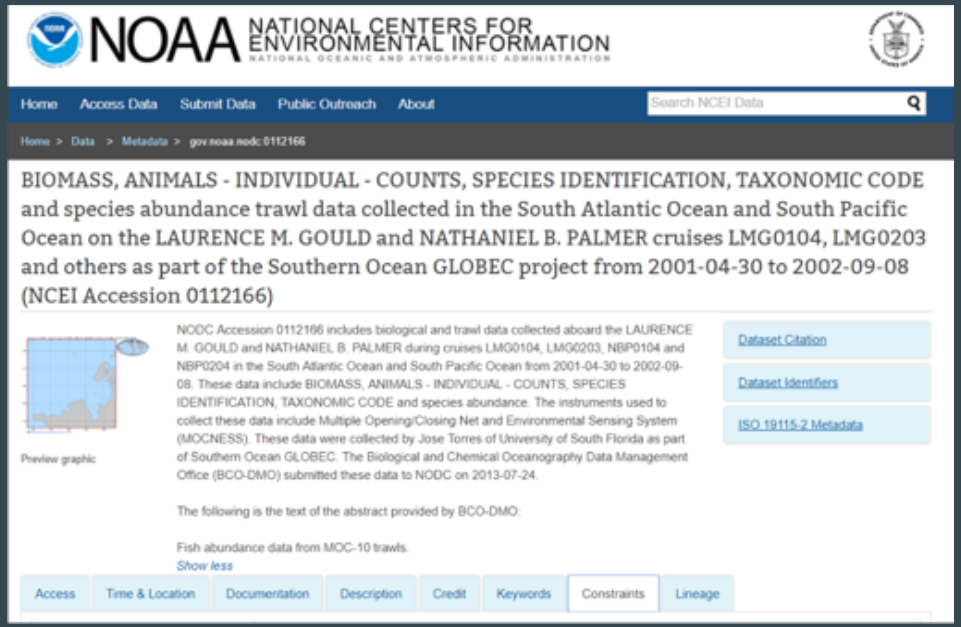
- BCO-DMO data holdings are freely accessible to the public
- No login or account creation needed
- Discoverable via text and geospatial search interfaces

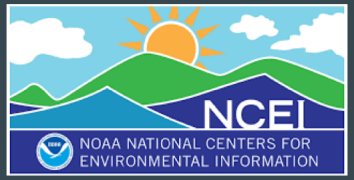


Discovery & Access

Preservation

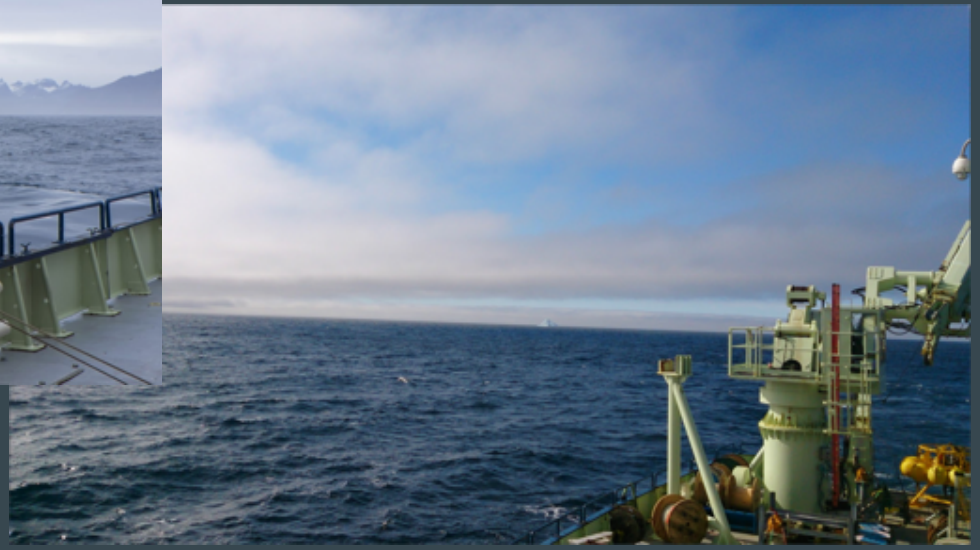
Once a project's data and metadata are published online at BCO-DMO, they are then submitted to an appropriate national data center for long-term preservation, e.g., the National Centers for Environmental Information (NCEI).





Successful Cruise Data Management...

Acquisition



Cruise Planning & Reporting

Cruise Plan (pre-cruise)

Document describing:

- Ship name, cruise ID
- Participants & their roles
- Dates & Ports of call
- Scientific objectives & planned operations
- Proposed cruise track & station locations
- List of instrumentation on board
- Station sampling plan & allocation of water/samples

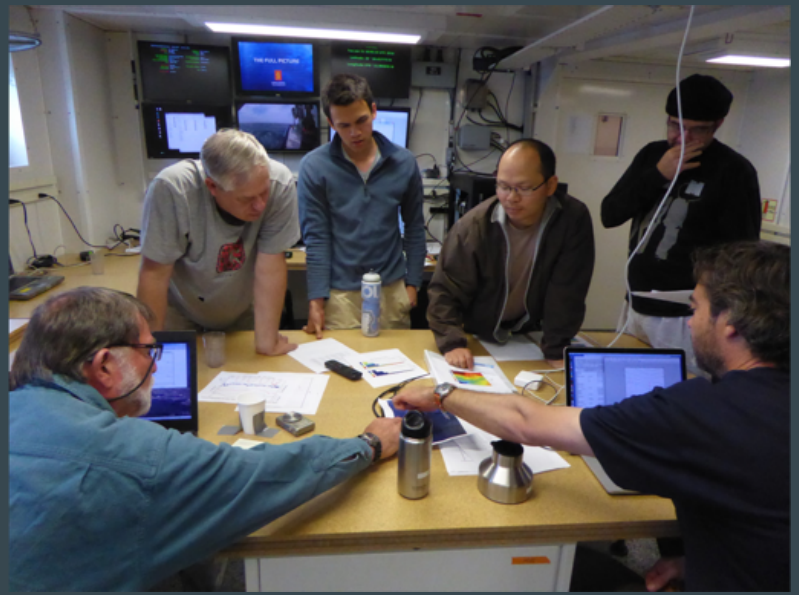
Cruise Report (post-cruise)

Document describing:

- Ship name, cruise ID
- Participants & their roles
- Actual dates & Ports of call
- Scientific objectives & operations accomplished
- Actual cruise track & station locations
- List of instrumentation used, description of problems encountered
- Preliminary results

Chief Scientist data responsibilities

- main point(s) of contact for BCO-DMO,
- submits event log, sample logs, cruise report, station list, hydrography, etc. to BCO-DMO,
- a list of expected datasets & the responsible PI -- very helpful!
- ** Make sure cruise participants know where to final version of logs & hydrography **



Event Logs

A chronological record of all scientific sampling events that happened during a cruise, wherein each sampling event is assigned a unique identifier. Recommended fields:

- Event #: unique to the cruise; perhaps unique across cruises for a multi-cruise project
- Instrument type/name/model
- Station #
- Cast # (if applicable)
- Date and time (specify UTC or local; if local, include time zone)
  Suggested format ISO: yyyy-mm-ddTHH:MM:SS
- Latitude and longitude
- Sampling depth (for ranges use depth_min and depth_max)
- Depth of the water
- Investigator
- Notes/comments

The event log allows investigators to integrate data from different sampling devices used during a cruise.

Event Logs

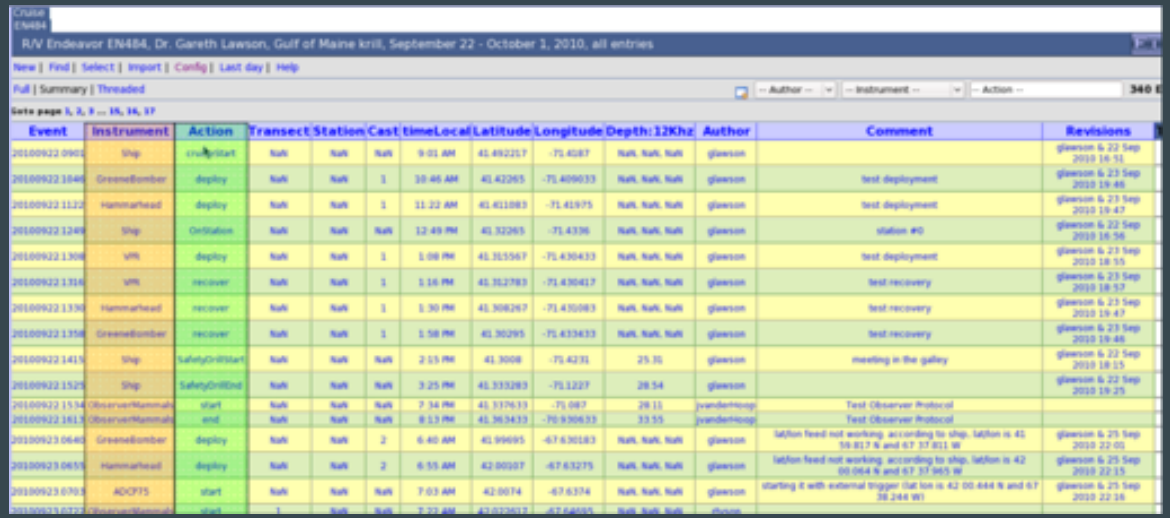
R2R
ROLLING DECK TO REPOSITORY

Scientific Sampling Event Log

The R2R Event Logger is installed on many UNOLS vessels.

The R2R Eventlogger is a program that creates a record of the scientific sampling events conducted during a cruise.

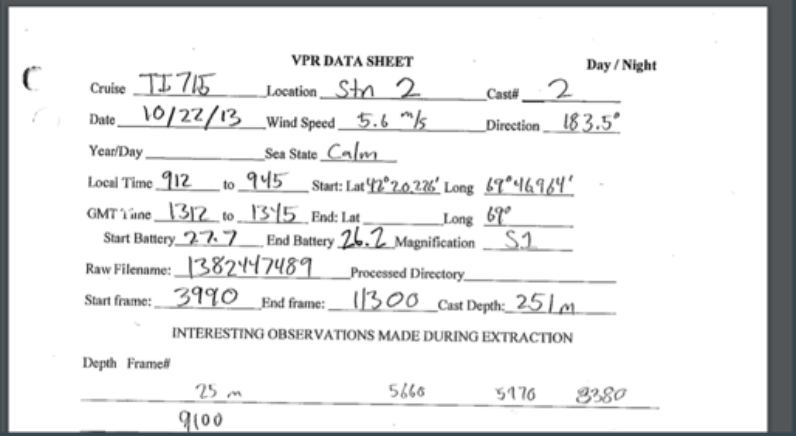
At the end of the cruise, the final event log can be exported as a plain text csv file.



See:
https://www.rvdata.us/about/event-log

Log Sheets

- Per sampling device
- Physical or digital
  - If hard copies, scan into PDFs and, ideally, transcribe into a spreadsheet



VPR DATA SHEET — Day / Night

Cruise TI715 Location Stn 2 Cast# 2

Date 10/22/13 Wind Speed 5.6 m/s Direction 183.5°

Year/Day ____ Sea State Calm

Local Time 912 to 945 Start: Lat 42°20.226' Long 69°46.964'

GMT Time 1312 to 1315 End: Lat ____ Long 69°

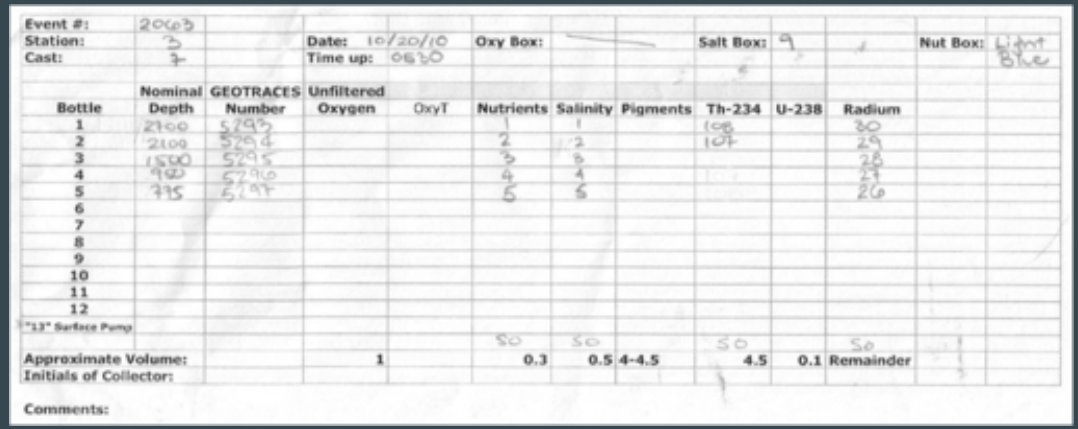
Start Battery 27.7 End Battery 26.2 Magnification 51

Raw Filename: 1382447489 Processed Directory ____

Start frame: 3990 End frame: 11300 Cast Depth: 25 m

INTERESTING OBSERVATIONS MADE DURING EXTRACTION

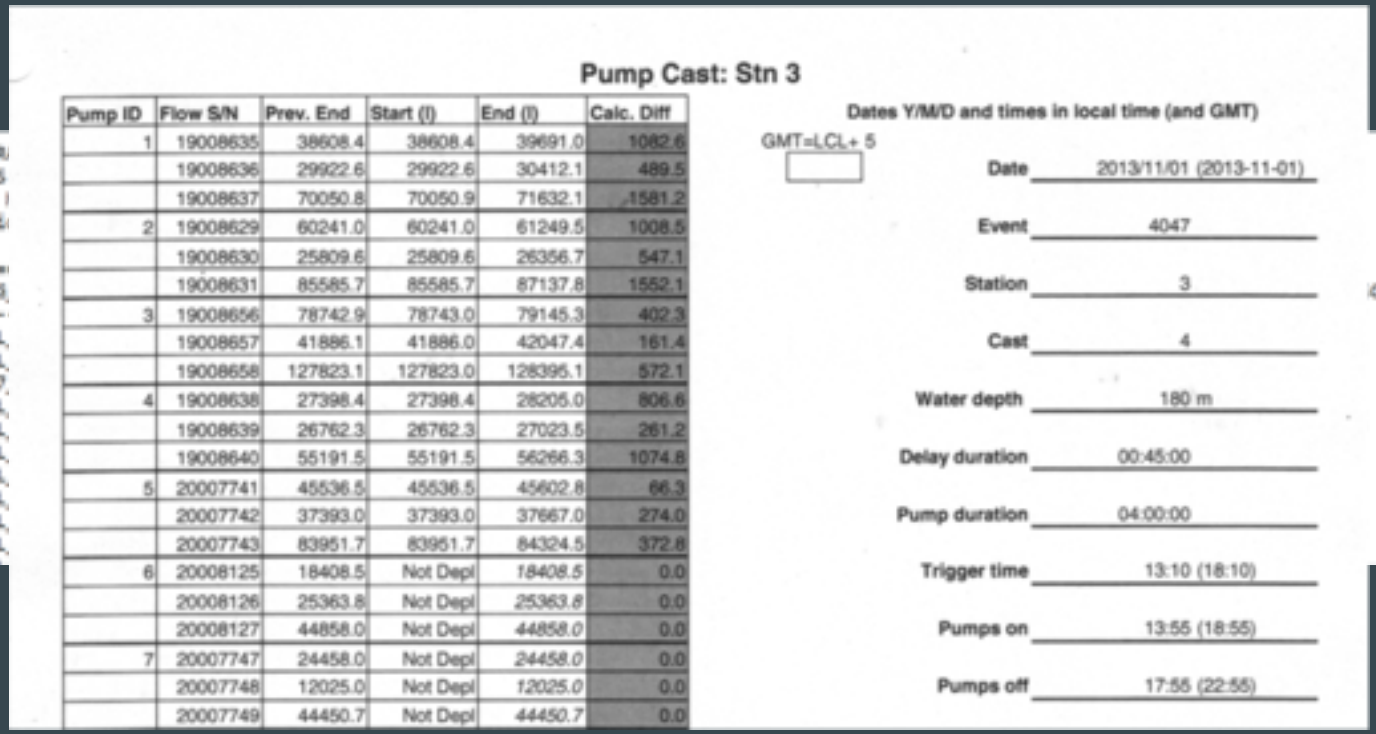
Depth	Frame#			
25 m	5660	5976	8380	9100



Event #:	20603											
Station:	3		Date:	10/20/10	Oxy Box:			Salt Box:	9		Nut Box:	Light Blue
Cast:	7		Time up:	0630								
Bottle	Nominal Depth	GEOTRACES Number	Unfiltered Oxygen	OxyT	Nutrients	Salinity	Pigments	Th-234	U-238	Radium		
1	2100	5793			1	1		108		30		
2	2100	5264			2	2		107		29		
3	1500	5295			3	3				28		
4	900	5290			4	4				27		
5	395	5299			5	5				26		
6												
7												
8												
9												
10												
11												
12												
"13" Surface Pump					50	50		50		50		
Approximate Volume:			1		0.3	0.5	4-4.5	4.5	0.1	Remainder		
Initials of Collector:												

Comments:

Log Sheets



Pump Cast: Stn 3

Pump ID	Flow S/N	Prev. End	Start (l)	End (l)	Calc. Diff
1	19008635	38608.4	38608.4	39691.0	1082.6
	19008636	29922.6	29922.6	30412.1	489.5
	19008637	70050.8	70050.9	71632.1	1581.2
2	19008629	60241.0	60241.0	61249.5	1008.5
	19008630	25809.6	25809.6	26356.7	547.1
	19008631	85585.7	85585.7	87137.8	1552.1
3	19008656	78742.9	78743.0	79145.3	402.3
	19008657	41886.1	41886.0	42047.4	161.4
	19008658	127823.1	127823.0	128395.1	572.1
4	19008638	27398.4	27398.4	28205.0	806.6
	19008639	26762.3	26762.3	27023.5	261.2
	19008640	55191.5	55191.5	56266.3	1074.8
5	20007741	45536.5	45536.5	45602.8	66.3
	20007742	37393.0	37393.0	37667.0	274.0
	20007743	83951.7	83951.7	84324.5	372.8
6	20008125	18408.5	Not Depl	*18408.5*	0.0
	20008126	25363.8	Not Depl	*25363.8*	0.0
	20008127	44858.0	Not Depl	*44858.0*	0.0
7	20007747	24458.0	Not Depl	*24458.0*	0.0
	20007748	12025.0	Not Depl	*12025.0*	0.0
	20007749	44450.7	Not Depl	*44450.7*	0.0

Dates Y/M/D and times in local time (and GMT)

GMT=LCL+ 5

Date 2013/11/01 (2013-11-01)

Event 4047

Station 3

Cast 4

Water depth 180 m

Delay duration 00:45:00

Pump duration 04:00:00

Trigger time 13:10 (18:10)

Pumps on 13:55 (18:55)

Pumps off 17:55 (22:55)

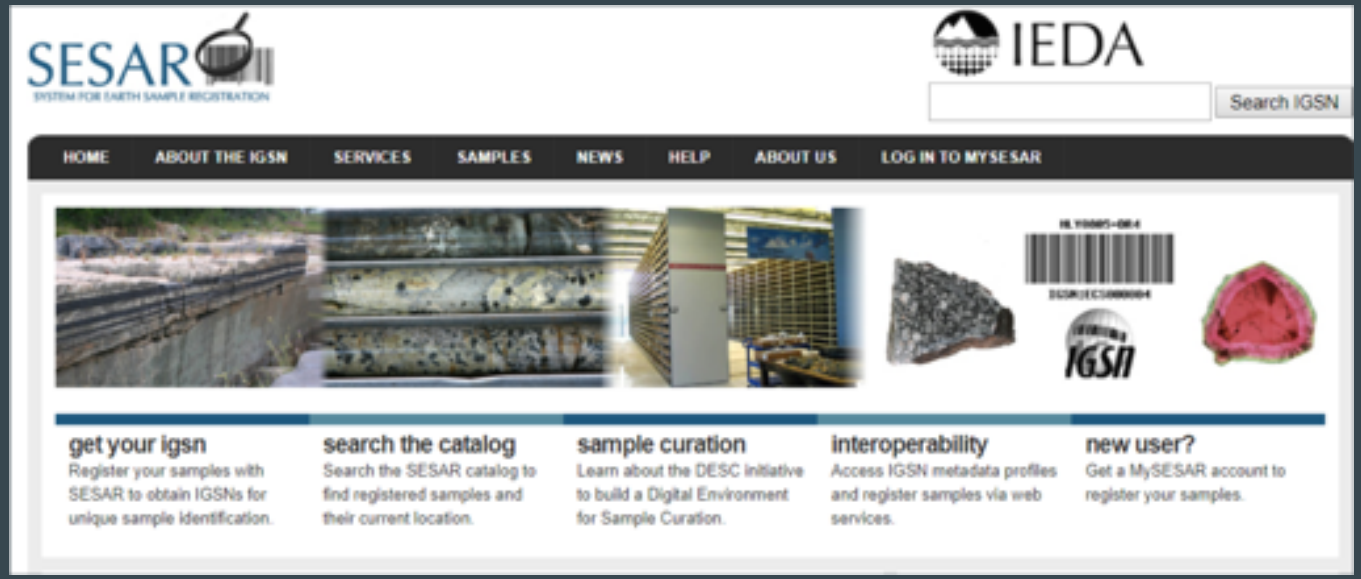
CRIPTION SCANNED_PUMP_LOG

4002_Hcl_Rinse2_2c.pdf
4013_Hcl_1_3c.pdf
4019_Be7_1_7s.pdf
4021_Hcl_1_9c.pdf
4023_Hcl_1_9s.pdf
4024_Hcl_1_12c.pdf
4024_Hcl_1_12s.pdf
4037_Hcl_2_3c.pdf
4047_Hcl_3_4c.pdf
4056_Hcl_4_3c.pdf

International Geo Sample Numbers (IGSNs)

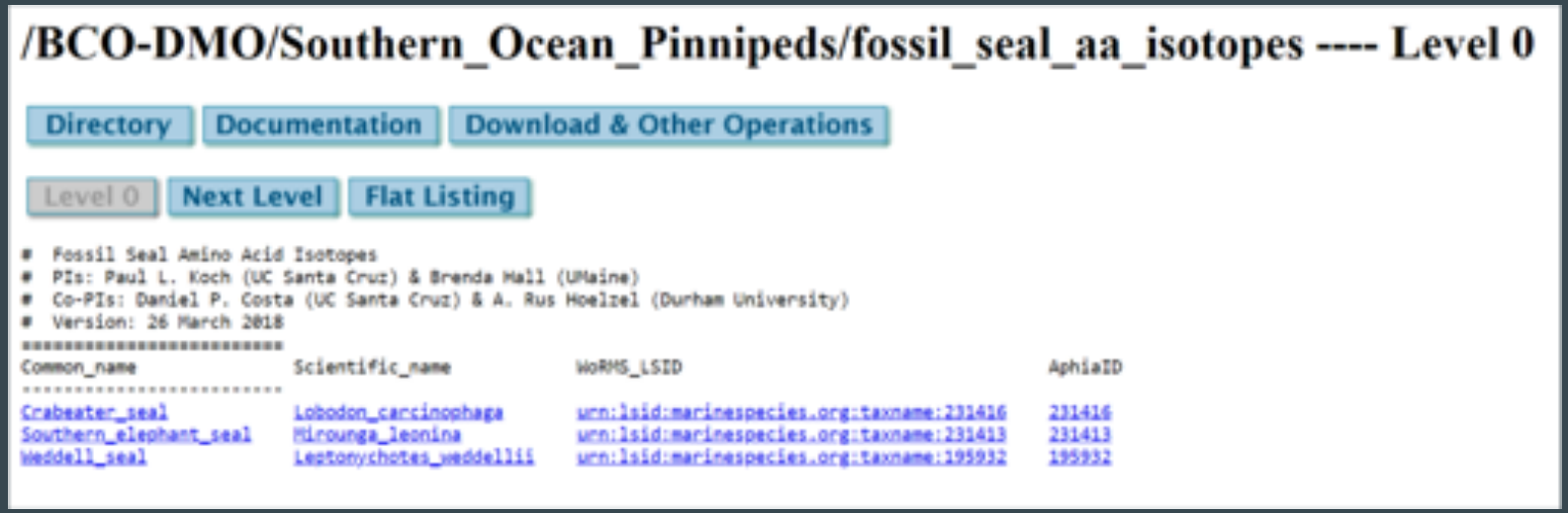
The IGSN is a persistent unique identifier for physical samples and specimens that eliminates the problems associated with the ambiguous naming of samples. In the U.S., you can obtain IGSNs using the System for Earth Sample Registration (SESAR) at IEDA Data Facility.

See https://igsn.github.io/overview/ and http://www.geosamples.org/



Taxonomy - WoRMs

The World Register of Marine Species (WoRMs) provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy.



/BCO-DMO/Southern_Ocean_Pinnipeds/fossil_seal_aa_isotopes ---- Level 0

Directory | Documentation | Download & Other Operations

Level 0 | Next Level | Flat Listing

```
#  Fossil Seal Amino Acid Isotopes
#  PIs: Paul L. Koch (UC Santa Cruz) & Brenda Hall (UMaine)
#  Co-PIs: Daniel P. Costa (UC Santa Cruz) & A. Rus Hoelzel (Durham University)
#  Version: 26 March 2018
```

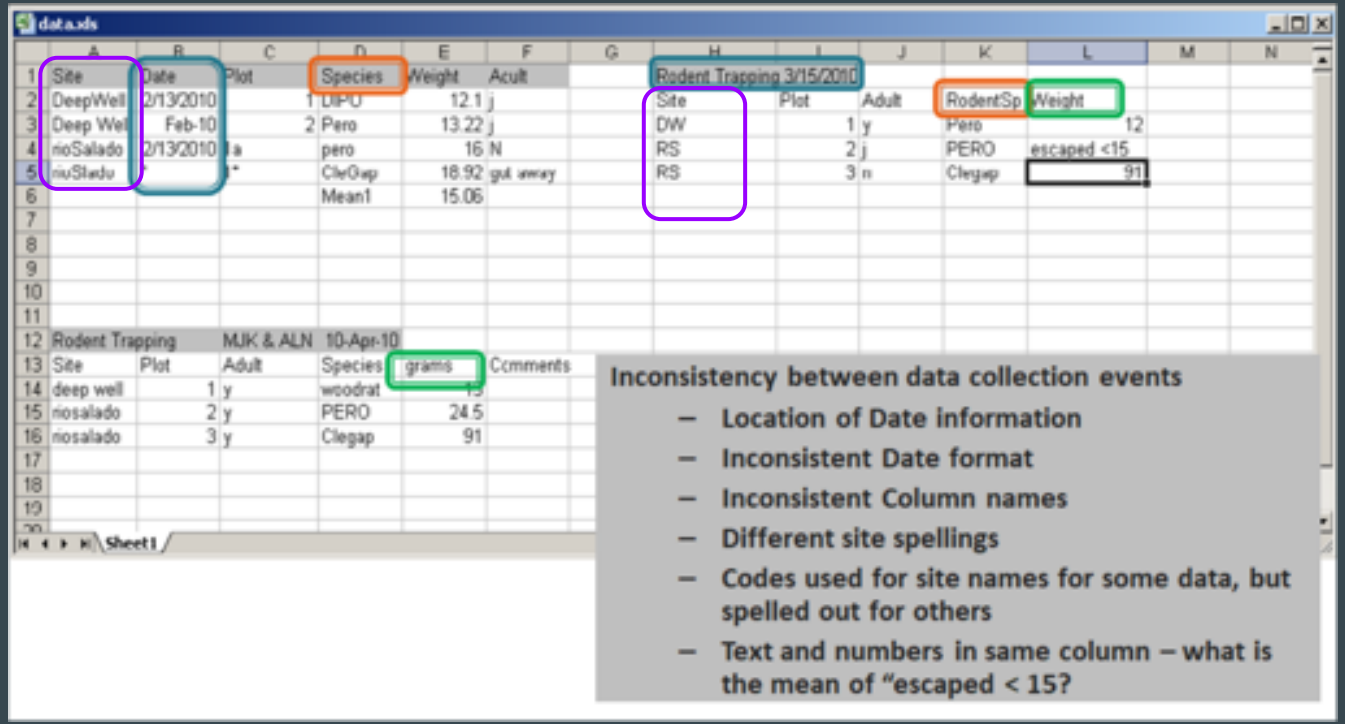
Common_name	Scientific_name	WoRMS_LSID	AphiaID
Crabeater_seal	Lobodon_carcinophaga	urn:lsid:marinespecies.org:taxname:231416	231416
Southern_elephant_seal	Mirounga_leonina	urn:lsid:marinespecies.org:taxname:231413	231413
Weddell_seal	Leptonychotes_weddellii	urn:lsid:marinespecies.org:taxname:195932	195932

We recommend checking species names in WoRMs and including identifiers in your data when possible.

Data Preparation Tips

- Use consistent formatting!
  - For site/station names/numbers, dates, times, position, column names (e.g. "EVENT" vs "event" vs "event_number" vs "Event Number")

Example: Poor Data Entry

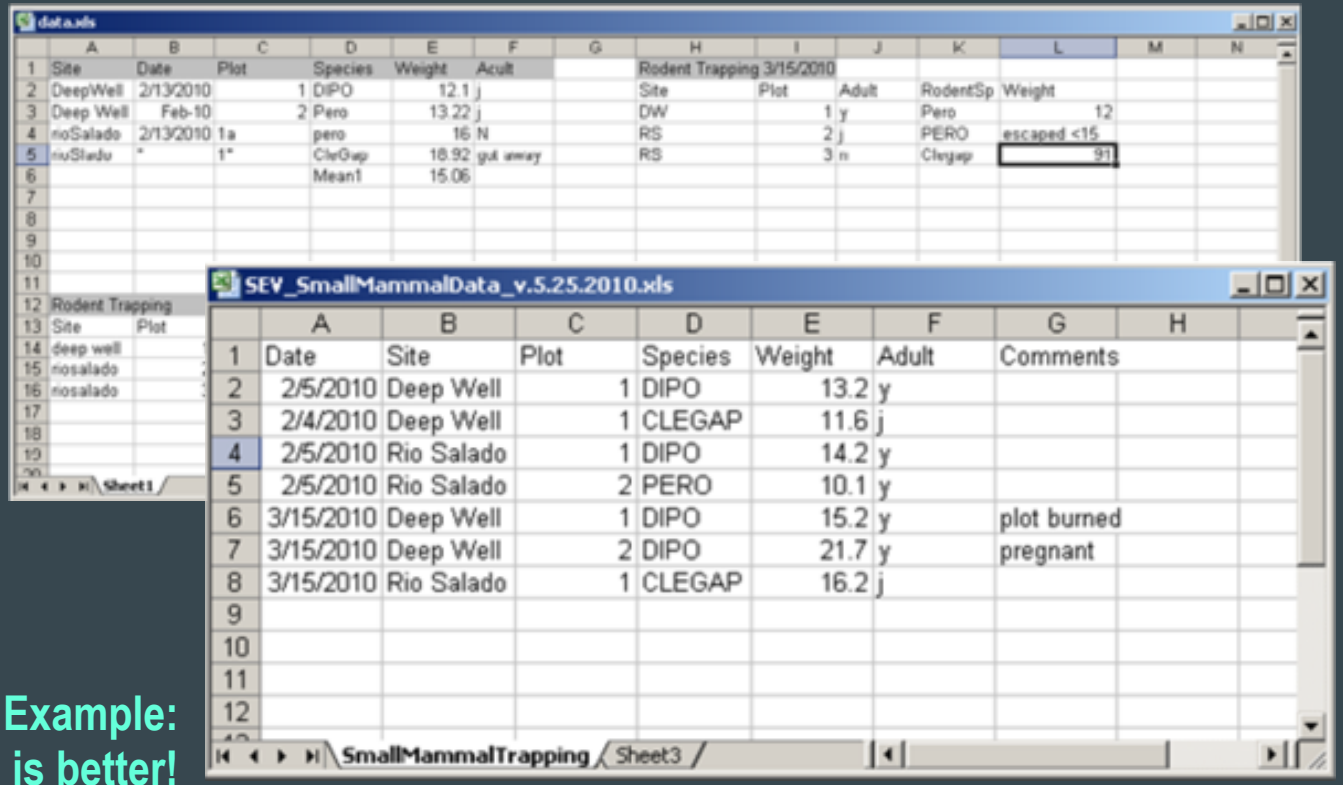


Data Preparation Tips

Columns of data are consistent: only numbers, dates, or text

Consistent names, codes, formats (date) used in each column.

Data are all in one table, which is much easier for a statistical program to work with than multiple small tables which each require human intervention.



Example:
This is better!

Data Formatting Best Practices

- Create descriptive column names without spaces or special characters. Use underscores instead of symbols.
  - Temp 30 meters → Temp_30_m
  - Species Code → species_code
- Avoid using numbers at the beginning of a column name (some programs have trouble with this)
- Use descriptive file names.
  - A file named PIV_E_gracilis_20180524.csv provides useful information about the data (type = Particle Image Velocimetry, species = E. gracilis, date= May 24, 2018).
  - Consistent formatting of file names allows for sorting and organizing of the files → keep in mind for images and video.
  - CTD files for example: cruiseID_station_cast.csv (e.g. KM1104_12_1.csv)

Data Formatting Best Practices

- Missing Data:
  - Again, be consistent!
  - Blank cells have no meaning...use "nd" or "NaN" (-999)
  - Bear in mind that "0" has meaning (0 = measured and not found as opposed to "not measured")
- Round data to appropriate number of decimal places
- Document all codes and quality flag definitions in the metadata
- Don't rely on Excel formatting to convey meaning (e.g. colored cells)

Any/all data: include lat/lon, date, and time whenever possible.

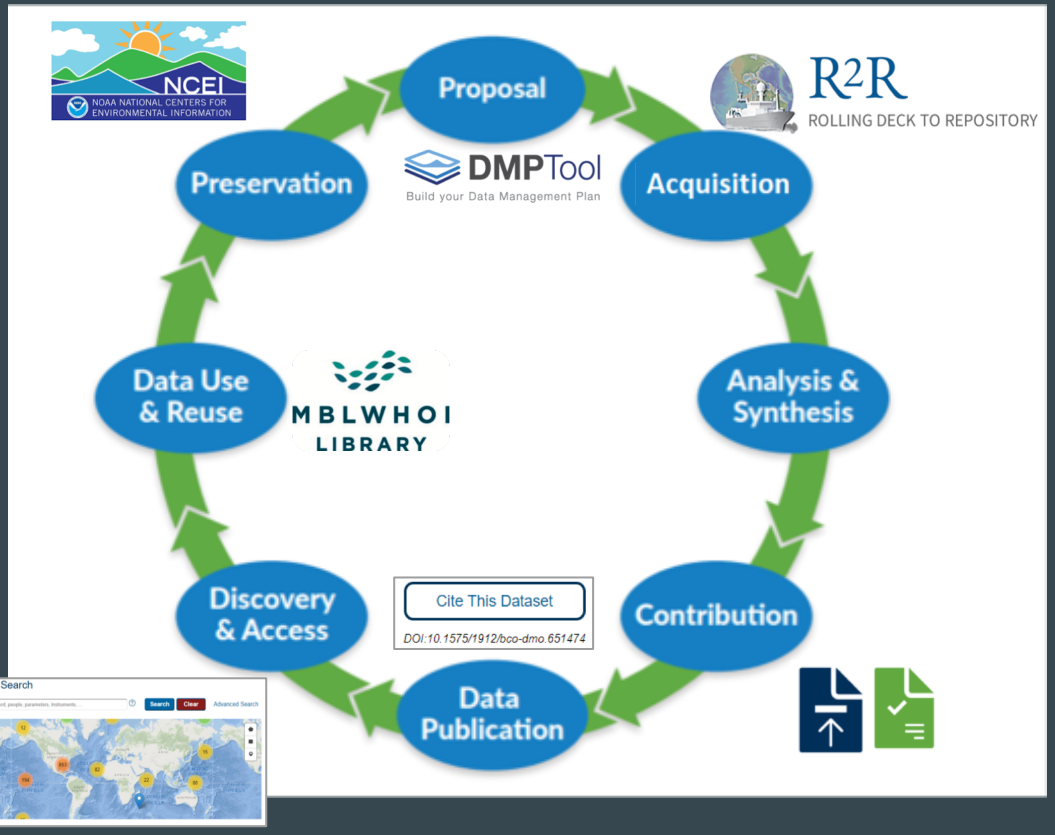
Quality Flags

You can incorporate data quality and uncertainty into your data and/or metadata.

e.g., IODE quality flags. See: https://www.iode.org/mg54_3

Value	Primary-level flag short name	Definition
1	Good	Passed documented required QC tests
2	Not evaluated, not available or unknown	Used for data when no QC test performed or the information on quality is not available
3	Questionable/suspect	Failed non-critical documented metric or subjective test(s)
4	Bad	Failed critical documented QC test(s) or as assigned by the data provider
9	Missing data	Used as place holder when data are missing

ODV, WOCE, other quality flag systems → use whatever is appropriate for your data/your community.


NCEI
NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
R2R
ROLLING DECK TO REPOSITORY
DMPTool
Build your Data Management Plan
Proposal
Acquisition
Analysis & Synthesis
Contribution
Data Publication
Discovery & Access
Data Use & Reuse
Preservation
MBLWHOI LIBRARY
Cite This Dataset
DOI:10.1575/1912/bco-dmo.651474
Dataset Search
Search
Clear
Advanced Search

Resources

- BCO-DMO: https://www.bco-dmo.org/resources (DMP Template, BCO-DMO Quick Guide)
- OCE Data Policy: https://www.nsf.gov/pubs/2017/nsf17037/nsf17037.jsp
- DMPTool: https://dmptool.org/
- R2R: https://www.rvdata.us/
- NCBI - how to submit: https://www.ncbi.nlm.nih.gov/guide/howto/submit-sequence-data/
- World Register of Marine Species: http://www.marinespecies.org/index.php
- IGSN: http://www.geosamples.org/
- DataONE Data Management Hub: https://dataoneorg.github.io/Education/
- Data Management Short Course: http://commons.esipfed.org/datamanagementshortcourse

Questions?

srauch@whoi.edu -- info@bco-dmo.org -- www.bcodmo.org