Fitting Square Pegs into a Round Hole

Curating Heterogeneous Oceanographic Data at BCO-DMO

Karen Soenen

BCO-DMO Team: Danie Kinkade, Adam Shepherd, Mak Saito, Dana Gerlach, Lynne Merchant, Sawyer Newman, Shannon Rauch & Amber York

> Ocean Sciences Meeting - February 22, 2023 Booth 508







18 Years of Oceanographic Data







Sharing is Caring (but it's hard!)

indable: Data are linked to descriptive persistent metadata.



Ccessible: Data and metadata are open, free, and machine

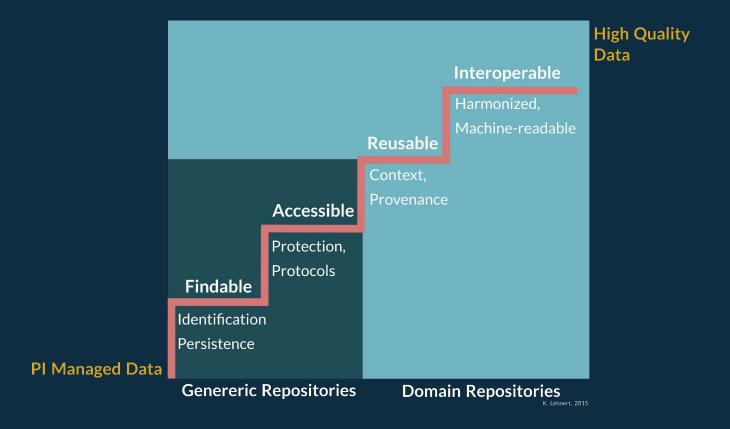


Data and metadata are standardized and use nteroperable: vocabularies. Data points to related metadata.



eusable: Metadata are rich, and employ usage licenses, provenance, and community standards.

Sharing is Caring (but it's hard!)

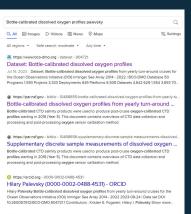


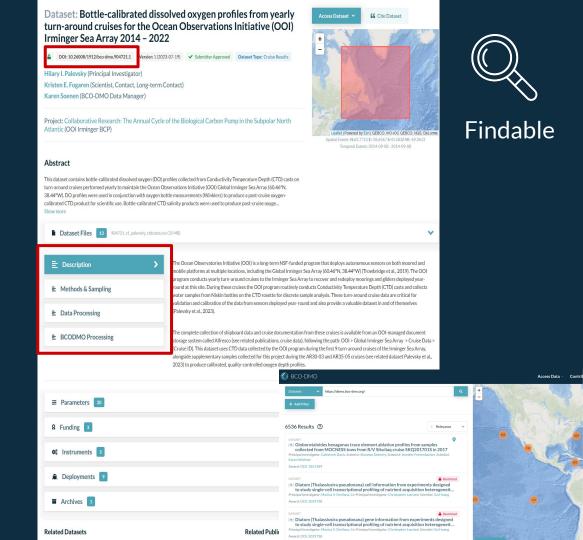
Can this dataset exist on its own and be reused?

Hypotheses come and go, but data remain.

https://www.bco-dmo.org/dataset/904721

- Unique ID (DOI)
- Rich Metadata
- Indexed in searchable resource





- (Meta) data are retrievable
- Embargo, until deadline set by funder requirements.

Dataset: Bottle-calibrated dissolved oxygen profiles from yearly turn-around cruises for the Ocean Observations Initiative (OOI) Irminger Sea Array 2014 - 2022



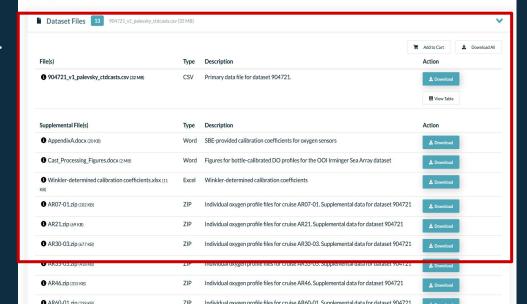
Atlantic (OOI Irminger BCP)





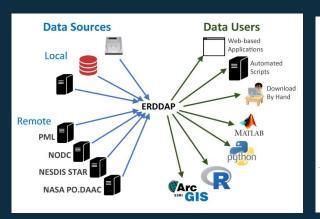
Abstract

This dataset contains bottle-calibrated dissolved oxygen (DO) profiles collected from Conductivity Temperature Depth (CTD) casts on turn-around cruises performed yearly to maintain the Ocean Observations Initiative (OOI) Global Irminger Sea Array (60.46°N. 38.44°W). DO profiles were used in conjunction with oxygen bottle measurements (Winklers) to produce a post-cruise oxygencalibrated CTD product for scientific use. Bottle-calibrated CTD salinity products were used to produce post-cruise oxyge... Show more

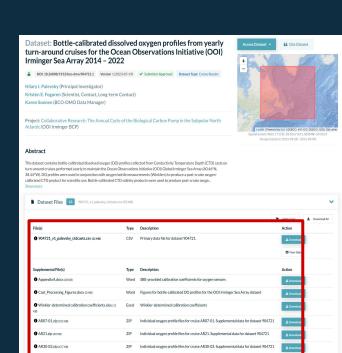


- (G) (G) (G)
- Interoperable

- Standards-compliant formats
- Machine readable <-> community uses







Individual oxygen profile files for cruise AR35-05. Supplemental data for dataset 904721

ZIP Individual oxygen profile files for cruise AR46. Supplemental data for dataset 904721

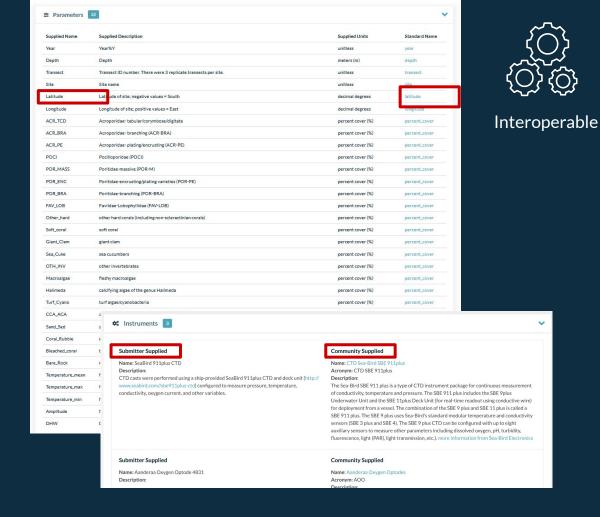
⊕ AR35-05.zip (418 KE)

♠ AR46 zip (315 kill)

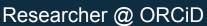
Adopting Controlled Vocabularies

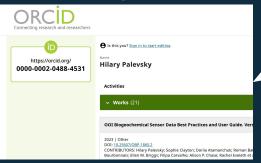
- Parameters
- Instruments





Connecting for Context





Related dataset (calibration) @ NCEI







Interoperable

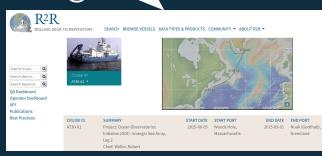
Award @ NSF



Publication



Cruise @ R2R





- Dataset QA/QC:
 - Adding location, date, time, unique identifiers, etc.
 - Harmonization (ISO Date Time)

Field Information			
Field	Units	Description	Data Type Format
ACR_BRA	percent cover (%)	Acroporidae- branching (ACR-BRA)	number
ACR_PE	percent cover (%)	Acroporidae- plating/encrusting (ACR-PE)	number
POCI	percent cover (%)	Pocilloporidae (POCI)	number
POR_MASS	percent cover (%)	Poritidae-massive (POR-M)	number
POR_ENC	percent cover (%)	Poritidae-encrusting/plating varieties (POR-PE)	number
POR BRA	percent cover (%)	Poritidae-branching (POR-BRA)	number

« Back to Metadata *Back to Metadata 918134_v1_benthic_comm_comp_heron_isl_2015-2020.csv

												918134_v1_benthic_comm_comp_heron_isl_2015-2020.csv								
Year	Depth	Transect	Site	Latitude	Longitude	ACR_TCD	ACR_BRA	ACR_PE	POCI	POR_MASS	POR_ENC	POR_BRA	FAV_LOB	Other_hard	Soft_coral	Giant_Clam	Sea_Cuke	OTH_INV	Macroalgae	Halimeda
2015	8.18	1	Harry's Bommie	-23,45977	151.9292	14.333	32.833	19.5	4	0	0	0	5.333	1.167	0.333	0	0.167	2.167	0.567	0
2015	8.18	2	Harry's Bommie	-23.45977	151.9292	8.848	9.791	4.206	3.086	1.563	0.089	0.693	2.259	0.780	0.441	0.341	0.379	0.846	9.164	3.993
2015	8.18	3	Harry's Bommie	-23.45977	151.9292	14.833	18	24	4.5	0.333	0	0	6.167	2.167	1.667	0	0.333	2	0	0
2015	6.11	1	Harry's Bommie	-23.45972	151.9295	17.333	42.833	7.833	2	1.333	0	0	3.833	1.5	0.5	0	0	0.333	1.533	0
2015	6.11	2	Harry's Bommie	-23,45972	151.9295	11.333	28.667	10.333	1.333	7.5	0	0	5.167	1.167	1.667	0	0.333	0.833	0.167	0
2015	6.11	3	Harry's Bommie	-23.45972	151.9295	23	19.333	13.833	3	1	0	0	6.167	4.833	0	0	0	1.333	0	0
2015	0.92	1	Reef Crest	-23.45887	151.9296	7.5	22.333	0	7.667	0.5	0	0.667	1.167	0	0	1	2	1.333	6	1.333
2015	0.92	2	Reef Crest	-23.45887	151.9296	1.167	49.333	2.5	4.5	5.833	0.667	0	3.333	0.167	0.5	0.5	0	0.833	6.833	4.667
2015	0.92	3	Reef Crest	-23.45887	151.9296	6.167	31	1.333	4.833	1.667	0	0	1.833	0	0	0	0.333	0.167	5.5	11
2015	0.71	1	Reef Flat	-23.45638	151.93	0.5	12.667	0	2.667	0.333	0	0.333	0.833	0	0	0	0.167	0.333	15.833	3
2015	0.71	2	Reef Flat	-23.45638	151.93	0.333	9.167	2.5	5	0	0	1.6	0.833	0.167	0	0	0.733	0.333	15	3.333

Declarative workflows





Declarative workflows in data processing

- Minimize error
- Streamline process
- Improved collaboration
- Software independent
- Provenance

22 submission files	904722_v1_palevsky_bottles & Connected to submission MicJnMqY\/vis9w7M						
•	√ Load [res2]	<u> </u>					
	Find and replace [res2], (Station): $^{(d)}$ \$ \rightarrow 00/1	<u> </u>					
4	Find and replace [res2], (Station): $^{(d(2))} \rightarrow 0$ 1	<u> </u>					
•	4 ▼ Load	<u> </u>					
	5 ▼ Concatenate data sources [ar35-05_012btl, ar35-05_013btl, ar35-05_011btl, ar30-03_021btl, ar30-03_015btl, ar30-03_010btl, ar30-03_004btl, ar3	<u></u> + ▶ ×					
	Split column [bottle_merged], (file_name) + Cruise, Cast	<u></u> + ▶ ×					
7	7 ▼ Find and replace [bottle_merged], (Cast]: btl.csv →	<u> </u>					
	Join data sources [res2, bottle_merged]	<u></u> + ▶ ×					
\$	Rename fields [bottle_merged] (Date (UTC)) + Date_UTC	<u></u> + ▶ ×					
,	Convert date [bottle_merged], (Date_UTC) + Date_UTC	<u></u> + ▶ ×					
•	**Reorder fields [bottle_merged]	<u></u> + ▶ ×					
,	** Set types [bottle_merged]	<u></u> + ▶ ×					
,	** Rename resource bottle_merged + 904722_v1_palevsky	<u></u> + ▶ ×					
	¹⁴ ▼ Dump final	<u></u> + ▶ ×					







Oceanographic data is diverse and complex

Disciplines

- Biological
- Chemical
- Biogeochemical
- Physical
- Geophysical

Collection Types

- In situ
- Laboratory
- Remotely sensed
- Synthetic/derived

Additional Challenges

- Variable organization
- Varying metadata
- Local parameter terms
- Emerging data types
- Distributed complementary info

Formats

- ASCII Text (tabular)
- Binary (e.g., NetCDF)
- Images
- Acoustics
- Application (e.g., Matlab)
- Links to other data

<u>Scale</u>

- Molecular to Megafaunal
- Local to Global
- Discrete to continuous /synoptic



Can this dataset exist on its own and be reused?

Hypotheses come and

but data remain.

(Ramón y Cajal, 1940)

Impact

search.dataone.org

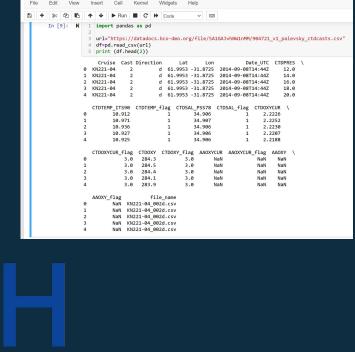
bco-dmo.org

benidmo ora

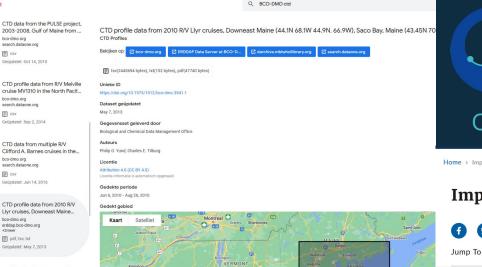
pdf, tsv, txt

search dataone org

Researcher **Research Community** Society



Jupyter Bottle_Data_Example Last Checkpoint: 16 minutes ago (autosaved)





Ocean InfoHub

Home > Improving Federal Programs Through Data and Evidence

Improving Federal Programs Through Data and Evidence







Challenges in Domain Curation

- It takes **time** to curate data -> often causes frustration of users
 - Long term data use vs paper publication needs
- Methodologies and instrumentation are evolving rapidly creating new data types of increasing size
- Templates & Data Management plans = structure. Once researchers adopt it, it
 makes it easy, but they have to get there first

info@bco-dmo.org





