

Semantics for Science

Part 1: WHY?

CYNTHIA L. CHANDLER, ADAM SHEPHERD, ROBERT A. ARKO
MATTHEW B. JONES AND DOUGLAS FILS

<http://bit.ly/S4S2017>

Semantics Symposium Bethesda, MD 10 January 2017



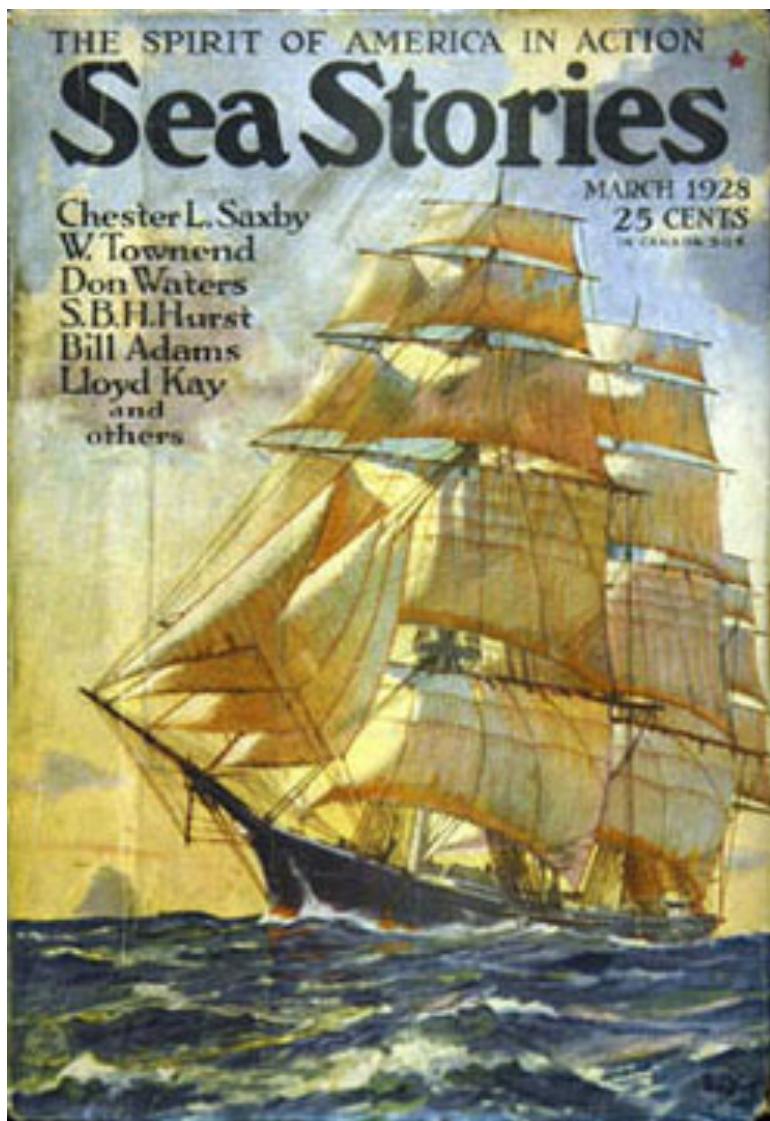
BCO-DMO
Biological & Chemical Oceanography Data Management Office

DataONE

CONSORTIUM FOR
Ocean Leadership

NSF

SEA STORY: PART 1



Google

North Atlantic Fishery



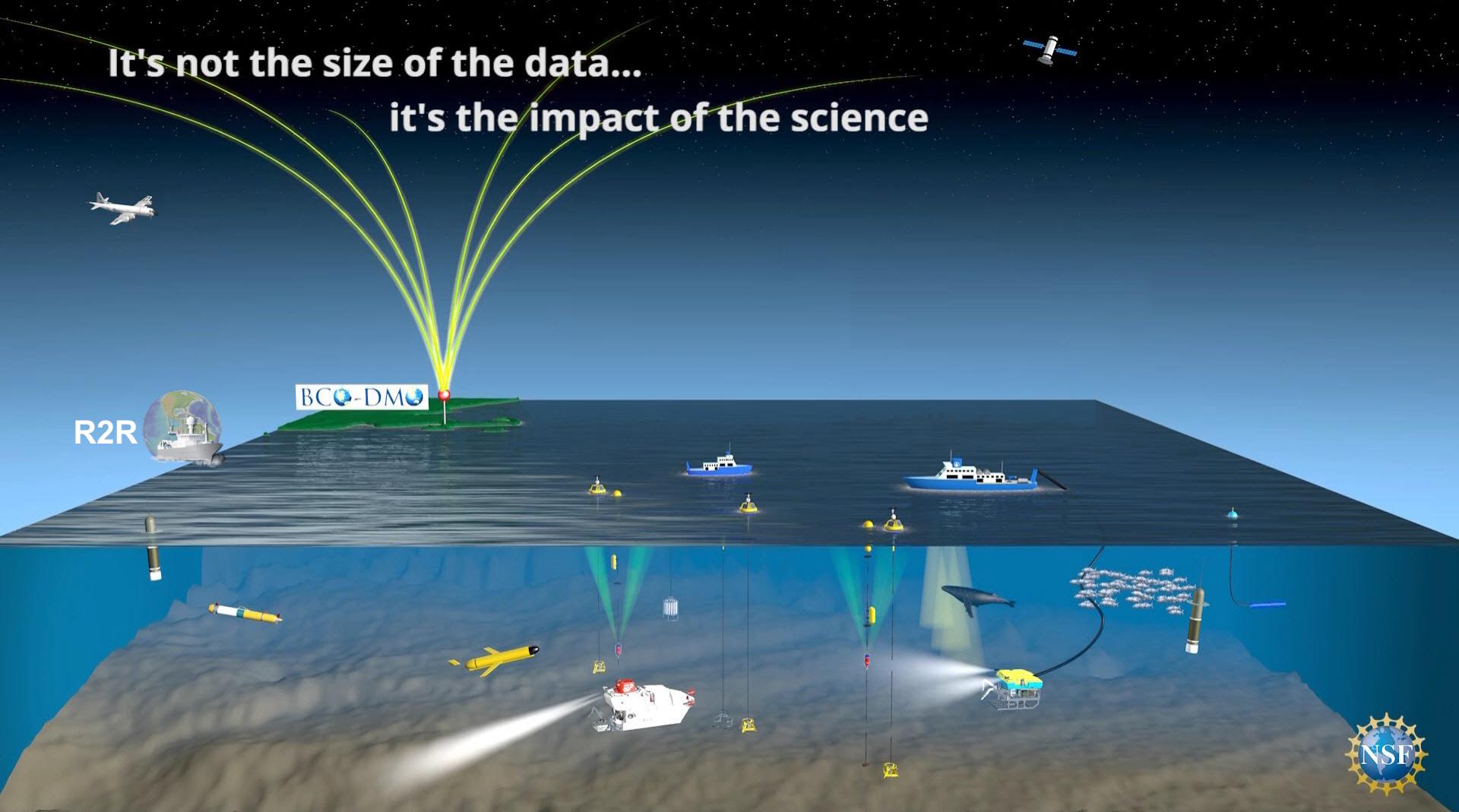
North Atlantic Cheap

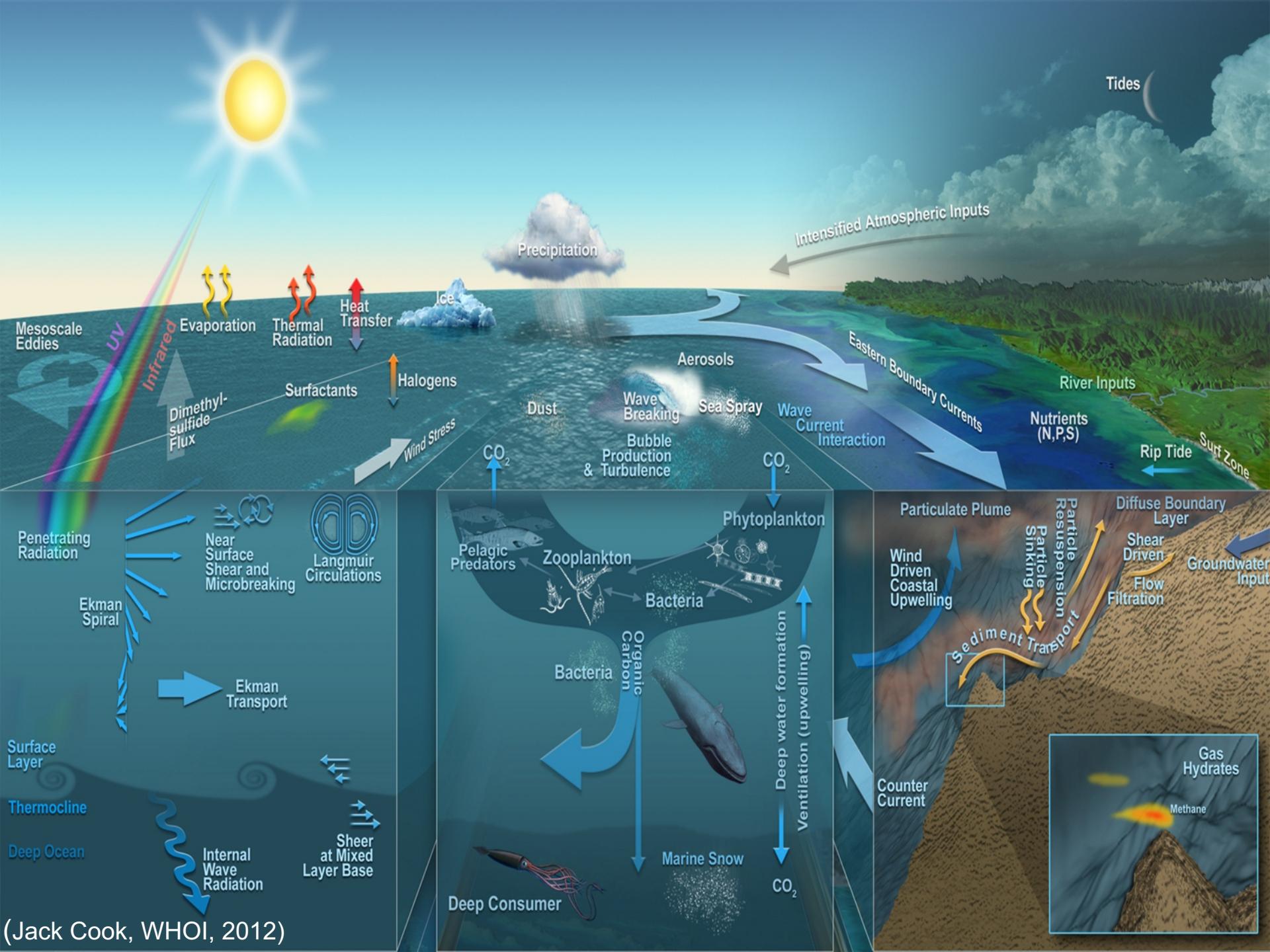
www.NexTag.com

Want Deals for North Atlantic? See NexTag Sellers' Lowest Price!

DISTRIBUTED DATA ... GLOBAL SCIENCE

**It's not the size of the data...
it's the impact of the science**





(Jack Cook, WHOI, 2012)

Distributed Related Resources

Funding Source

The screenshot shows the NSF website with a search bar at the top. Below it, a sidebar on the left lists various award categories like 'Search Awards' and 'Recent Awards'. The main content area displays an award abstract for 'Award Abstract #0752336 Oceanographic control and global distributions of subseafloor microbial life and activity'. It includes sections for 'NSF Org: OCE Division Of Ocean Sciences', 'Initial Amendment Date: March 21, 2008', 'Latest Amendment Date: March 21, 2008', 'Award Number: 0752336', 'Award Instrument: Standard Grant', 'Program Manager: David L. Garrison, OCE Division Of Ocean Sciences, GEO Directorate For Geosciences', 'Start Date: September 1, 2008', and 'End Date: August 31, 2013 (Estimated)'.

Field Expedition (e.g. cruise)

The screenshot shows the R2R website with a map of the North Pacific Ocean. A purple polygon indicates the cruise route. The map shows bathymetry and current flow patterns. Below the map, the text reads: 'Operator: Woods Hole Oceanographic Institution Vessel (retired): Knorr'. At the bottom, it says 'Cruise DOI: 10.7284/900496 #'. The page also includes a 'Catalog Status' section with 'On Service' status and a 'Catalog ID' of 60177, and a 'Archived Files' section with a file ID of 28917251 from June 23, 2016.

Data Sets

The screenshot shows the BCO-DMO website. The main header says 'Dataset: Subseafloor Microbial Cell Counts Deployment: KN195-03'. Below this, it states 'Microbial cell counts in sediment cores collected during KN195-03.' It lists the principal investigator as Dr. Steven L. D'Hondt (University of Rhode Island, URI-GSO). The dataset includes 36 programs, 687 projects, 2357 deployments, 8062 datasets, 408 instruments, 1372 parameters, 2093 people, 485 affiliations, 83 funding, and 1963 awards. The project description mentions 'Oceanographic control and global distributions of subseafloor microbial life and activity'.

Physical Samples

The screenshot shows the SESAR website with a search results page for 'Sample Search'. It lists 795 samples found, with a table showing columns for 'Object Type', 'Material Classification', 'Latitude', 'Longitude', and 'Location'. The table includes rows for various sediment samples from the KN195-03 cruise, such as KN195_S1_MC1 through KN195_S3A_JC2.

The screenshot shows the EarthChem library record for 'Dataset Title: Dissolved Inorganic Carbon measurements of marine sediment from Long Core Expedition KN195-03'. It lists authors (Sauvage, Justine; D'Hondt, Steven; Pockalny, Robert), dataset DOI (doi:10.1594/IEDA/100562), and date released/published (10/15/2015). The abstract describes the research objective of testing quantitative models for the magnitude and geographic distribution of subseafloor biomass and organic matter in the North Pacific Gyre. The citation is Sauvage, Justine; D'Hondt, Steven; Pockalny, Robert (2015): Dissolved inorganic carbon measurements of marine sediment from Long Core Expedition KN195-03. EarthChem Library. doi:10.1594/IEDA/100562.

Publications

The screenshot shows the Science journal website with a news article titled 'Aerobic Microbial Respiration in 86-Million-Year-Old Deep-Sea Red Clay'. The article is authored by Hans Roy¹, Jens Kallmeyer², Rishi Ram Adhikari², Robert Pockalny³, Bo Barker Jørgensen¹, Steven D'Hondt³. The abstract discusses the knowledge of the chemical composition of the porewater throughout the core and the potential for aerobic microbial respiration in the deep-sea sediments of the expedition. The shipboard geochemistry laboratory generated interstitial waters by squeezing 977 sediment samples. The authors note that the number of samples was much higher than the time needed to analyze them, so not all were analyzed for each dissolved constituent. They present here the shipboard porewater dissolved inorganic carbon concentration data and discuss the implications for the long-term preservation of the deep-sea environment. The article is published in Science, Vol. 336, Issue 6083, pp. 922-925, DOI: 10.1126/science.1219424.

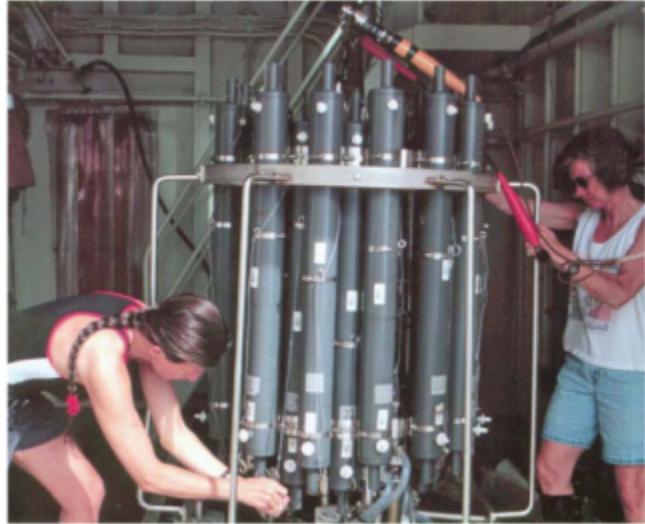
R2R, BCO-DMO, DATAONE

- R2R provides stewardship of routinely-acquired environmental sensor data from U.S. academic research vessels, and a catalog of cruises
- BCO-DMO works in partnership with NSF-funded researchers to help improve access to marine research data some of which is acquired during research cruises
- DataONE is a global network that federates content from many repositories including R2R and BCO-DMO



Repositories funded primarily by U.S. National Science Foundation
with additional support from NOAA, ONR, GBMF and SOI

FREE TEXT METADATA – NOT ENOUGH



Found we needed to disambiguate as much of the metadata content as possible.

Sufficient metadata is essential but plain text metadata is not enough. Semantic Web technologies help to improve discovery and access.

PERSISTENT IDENTIFIERS

Current practices in U.S. ocean science:

DOIs for Articles

DOIs for Datasets

DOIs for Documents

FundRef codes for Awards

IGSNs for Samples

ORCIDs for Persons

(Hanson, 2016; 10.1029/2016EO043183)



PERSISTENT IDENTIFIERS

Emerging practices in U.S. ocean science:

- GRID IDs for Organizations
- DOIs for Collections (package of objects)
- DOIs for Repositories (via RE3data.org)
- DOIs for Expeditions (via e.g. R2R, IODP)
- DOIs for Networks (via e.g. FDSN)

PROGRESS SO FAR . . .

Metadata, plus
Semantic markup, plus
Persistent identifiers . . .



PROGRESS SO FAR . . .

Metadata, plus
Semantic markup, plus
Persistent identifiers



Linking Related Resources via PIDs

Person @ORCID

Paper

A screenshot of a Science magazine article. The title is "Aerobic Microbial Respiration in 86-Million-Year-Old Deep-Sea Red Clay". The authors listed are Hans Roy^{1,*}, Jens Kallmeyer², Rishi Ram Adhikari², Robert Pockalny³, Bo Barker Jørgensen¹, Steven D'Hondt³. The journal is Science, Vol. 336, Issue 6083, 18 May 2012. DOI: 10.1126/science.1219424.

A screenshot of the BCO-DMO dataset deployment page for KN195-03. The dataset is titled "Dataset: Subseafloor Microbial Cell Counts". Deployment information includes "Deployment: KN195-03" and "Microbial cell counts in sediment cores collected during KN195-03". The principal investigator is Dr. Steven L. D'Hondt (University of Rhode Island, URI-GSO). The dataset is described as "Dissolved inorganic carbon measurements of marine sediment from Long Core Expedition". The dataset has a DOI of doi:10.1194/EDB/100562 and was released on 10/11/2012. A detailed abstract and citation are provided at the bottom.

Dataset @BCO-DMO

A screenshot of the ORCID profile for Jens Kallmeyer. It shows his ORCID ID (0000-0002-6440-1140) and other IDs (Loop profile: 30680, ResearcherID: I-3554-2012). His education is listed under Max-Planck-Institute for Marine Microbiology, Bremen, Bremen, Germany (2000-02 to 2003-09-30 (Biogeochemistry)).

Cruise @R2R

A screenshot of the Rolling Deck to Repository (R2R) cruise catalog for KN195-03. The catalog status shows 26 vessels, 6117 cruises, and 22917261 archived files. The map shows the cruise route in the North Pacific. The operator is Woods Hole Oceanographic Institution Vessel (retired): Knorr. The cruise DOI is 10.7284/90049.

A screenshot of the EarthChem library record for KN195-03. The dataset title is "Dissolved inorganic carbon measurements of marine sediment from Long Core Expedition". The dataset has a DOI of doi:10.1194/EDB/100562 and was released on 10/11/2012. The dataset is described as "The KN195-03 research cruise on board the R/V Knorr (chief scientist Steven D'Hondt) took place in January 2009 and had a primary objective to test and refine quantitative models for the magnitude and geographic distribution of subseafloor biomes and organic-fueled subseafloor respiration in the central North Pacific Gyre. Knowledge of the chemical composition of the porewater throughout the ocean is critical for understanding the dynamics of the subseafloor biosphere and its role in the global carbon cycle. The shipboard geochemistry laboratory generated interstitial waters by squeezing 997 individual samples, and by extracting an additional 153 solutions by filtration sampling. Due to the large number of samples, the shipboard geochemistry laboratory developed a rapid method to measure dissolved constituent. We present here the shipboard porewater dissolved inorganic carbon concentration dataset, measured by Marandis ARICA™ infrared system."

Dataset @ECL

A screenshot of the National Science Foundation award abstract for #0752336. The award title is "Oceanographic control and global distributions of subseafloor microbial life and activity". The award number is 0752336, and the program manager is David L. Garrison. The start date is September 1, 2008. The award abstract is available at https://nsfgeosamples.org/cruise_search.php?cruise=KN195-03&pp=All

NSF Org:	GCE	Division Of Ocean Sciences
Initial Amendment Date:	March 21, 2008	
Latest Amendment Date:	March 21, 2008	
Award Number:	0752336	
Award Instrument:	Standard Grant	
Program Manager:	David L. Garrison GCE Division of Ocean Sciences GEO Directorate for Geosciences	
Start Date:	September 1, 2008	

SESAR

Sample Search

Samples Found: 795												
View	Preview	1	2	3	4	5	6	7	All	Items per page:	All	Page 1 of 32
view		ISBN	Sample Name	Object Type	Material Classification	Latitude	Longitude	Location				
view		IE2230020	KN195_S1_Mc1	Core	Sediment	1.803485	-86.189645					
view		IE2230020	KN195_S1_Gc1	Core	Sediment	1.803485	-86.189645					
view		IE2230020	KN195_S1_Lc1	Core	Sediment	1.803485	-86.189645					
view		IE2230020	KN195_S1_Lc2	Core	Sediment	1.803485	-86.189645					
view		IE2230020	KN195_S2_Mc7	Core	Sediment	-4.23900333333333	-92.968245					
view		IE2230020	KN195_S2_Gc1	Core	Sediment	-4.23900333333333	-92.968245					
view		IE2230020	KN195_S2_Gc2	Core	Sediment	-4.23900333333333	-92.968245					
view		IE2230020	KN195_S3_Gc1	Core	Sediment	-0.0110916666666667	-104.352478333333					
view		IE2230020	KN195_S3_Am1	Core	Sediment	-0.044835	-105.425105					
view		IE2230020	KN195_S3_Gc1	Core	Sediment	-0.044835	-105.425105					
view		IE2230020	KN195_S3_Lc1	Core	Sediment	-0.044835	-105.425105					
view		IE2230020	KN195_S3_Lc2	Core	Sediment	-0.044835	-105.425105					

Samples @SESAR



Linking Additional Resources

educational history



related articles

A screenshot of the PublGrid search interface showing results for a query related to deep-sea microbiology.

A collage of scientific web pages. On the left is a Science journal article titled "Aerobic Microbial Respiration in 86-Million-Year-Old Deep-Sea Red Clay". In the center is a BC-DMC dataset page for "Subseafloor Monolith Cell Counts". To the right are several other research catalog and database pages, including one from the National Science Foundation (NSF) and another from IEDV.



instrumentation details



research vessel information

related research

A screenshot of the National Science Foundation (NSF) website featuring a news article about a newly discovered hydrothermal vent life form.

photos of cores



GEOLINK ... AND BEYOND

- Use ontology design patterns
- Controlled vocabulary term URIs (PIDs)
- Promote ORCIDs (person) and DOIs
- Publish more resources as Linked Data from other repositories
- BCO-DMO Linked Data <ISO 19115, term URIs, links to other term URIs>

GeoLink ontology patterns: <http://schema.geolink.org/>

LINKED DATA + ISO 19115

BCO-DMO publishes metadata out as Linked Data.

- RDF with persistent identifiers (e.g. ORCID for person, and DOI for cruise)
- Link to the full ISO 19115 record

<http://lod.bco-dmo.org/id/dataset/3045.rdf>

```
<rdf:Description rdf:about=
  "http://lod.bco-dmo.org/id/dataset/3045#iso">
<rdfs:seeAlso rdf:datatype=
  "http://www.w3.org/2001/XMLSchema#anyURI">
  http://www.bco-dmo.org/dataset/3045/iso
</rdfs:seeAlso>
```

THANK YOU



EARTH CUBE
TRANSFORMING GEOSCIENCES RESEARCH



GeoLink Partners:

Representing: Marine Sciences, Library Sciences, and Computer Sciences

Woods Hole Oceanographic Institution

- Cynthia Chandler
- Adam Shepherd
- Peter Wiebe
- BCO-DMO staff members

Lamont-Doherty Earth Observatory

- Robert Arko, Peng Ji
- Suzanne Carbotte, Kerstin Lehnert

MBLWHOI Library

- Lisa Raymond, Audrey Mickle

Ocean Leadership: Doug Fils

Marymount University

- Tom Narock

University of CA, Santa Barbara

- Matt Jones (NCEAS, DataONE)
- Yingji Hu, Bryce Mecum
- Krzysztof Janowicz
- Mark Schildhauer

Wright State University

- Pascal Hitzler
- Michelle Cheatham
- Adila Krisnadhi

AUDIENCE PARTICIPATION



CHALLENGES

What other ways are people using semantics to meet information management challenges?

Provenance?

CONSIDERATION

Should we publish PIDs for controlled vocabulary terms?

e.g. classes of instruments



sameAs



EXTRA SLIDES – PID SEMANTICS

CONSIDERATION

What other PID systems are in use?

CHALLENGE

DOIs issued for multiple copies – different instances of same content

e.g. same data at many repositories
copies of the same data set could appear in many repositories, with different metadata as determined by the repository practices, with each repository assigning a new DOI; in many cases the repositories may not be aware that the data have already been published by another repository

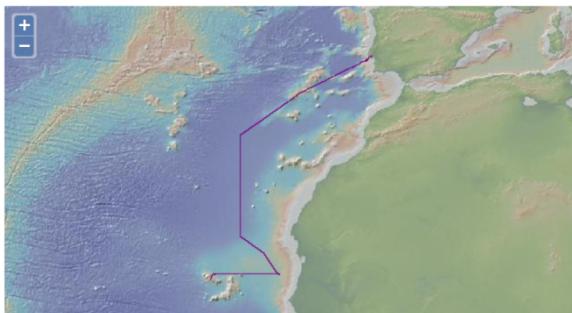
CONCERN

DOI assignment practices

e.g. assign DOI for an event as well as collection of results from an event; is there a recommended practice for citation?

Assign a DOI to the metadata record for a sampling event (e.g. cruise)

Cruise Catalog: KN199-04



Operator: Woods Hole Oceanographic Institution
Vessel (retired): Knorr

Cruise DOI: 10.7284/900522

Assign a DOI to the collection of all data sets and documents at a repository that are associated with an event

CCHDO Home Find Data Submit Data Information Search

Hydrographic Cruise: 316N20101015

Dataset

Files in the Dataset are the data for this cruise. They are updated when new data are submitted or as needed.

Download Entire Dataset Submit Data For This Cruise

ctd

- exchange: 316N20101015_ctd.zip (958.5 kB)
- wwp_ncetcd: 316N20101015_nc_ctd.zip (1.1 MB)

documentation

- pdf: 316N20101015_dc.pdf (6.6 MB)

Unmerged Data as Received

Files listed here are updates to the dataset which have not been processed yet, they may not be well formatted. Data files listed here usually contain the most up to date versions of the data for specific parameters. If you are unsure of which files to use, stick to the Dataset files above.

These files are not yet in the Dataset.

Filename (Download)	Size	Date Submitted
README_1410.txt	11.4 kB	2014-11-12
GT10_NAZT_Bottle_Data.zip	86.2 kB	2014-11-12
gt10_CruiseReport.zip	1.5 MB	2013-08-07

CHALLENGE

"upstream/downstream" issue

Are there recommended practices for encoding provenance using PIDs.

Datasets pointing to parent event.

Articles reference datasets.

ORCID points to projects.

CHALLENGE

Granularity of DOI assignment: DataCite expects DOI PIDs to represent a data set (e.g. with a title), whereas we need PIDs both for that and for components of data sets such as individual files, granules, and records.

e.g. use UUIDs for internal components of data sets, and assign the DOI to the whole data set.

There could be good reason for DOIs to be assigned to more granular components. In general, the PID community has not embraced the need for PIDs at multiple levels of the data hierarchy (cell, record, file, metadata, package).

CHALLENGE

Person IDs: ORCIDs working great for extant people, still challenged on how to assign person identifiers for all of our historical data records (e.g. people who have moved on to other fields)

CONCERN

If a dataset or document isn't published with a DOI, does it exist?

Will non-DOI'd content be discoverable by emerging networks like CrossCite tool, Scholix, etc?

CONCERN

A DOI is a DOI ... or are CrossRef and DataCite data set DOIs different ?

One of these things is not like the other ...
but is there a lossless crosswalk between their metadata schemas?



THANK YOU



Research funded by: NSF OCE-1435578 (BCO-DMO), ICER-1440114 (R2R),
OCE-1447797 (EarthCube GeoLink)

BCO-DMO: <http://bco-dmo.org/>
R2R: <http://rvdata.us/>

DataONE: <http://dataone.org/>
IODP: <http://iodp.org/>