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

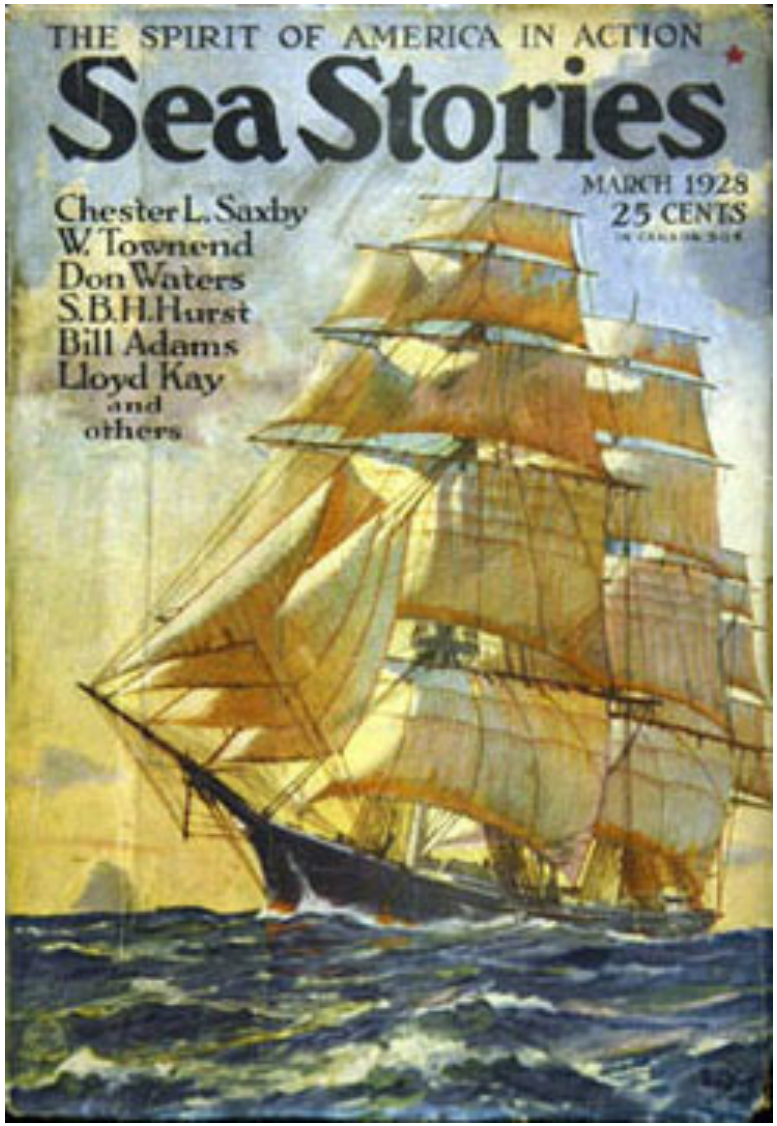
BCO-DMO
Biological & Chemical Oceanography Data Management Office

DataONE

CONSORTIUM FOR Ocean Leadership

NSF

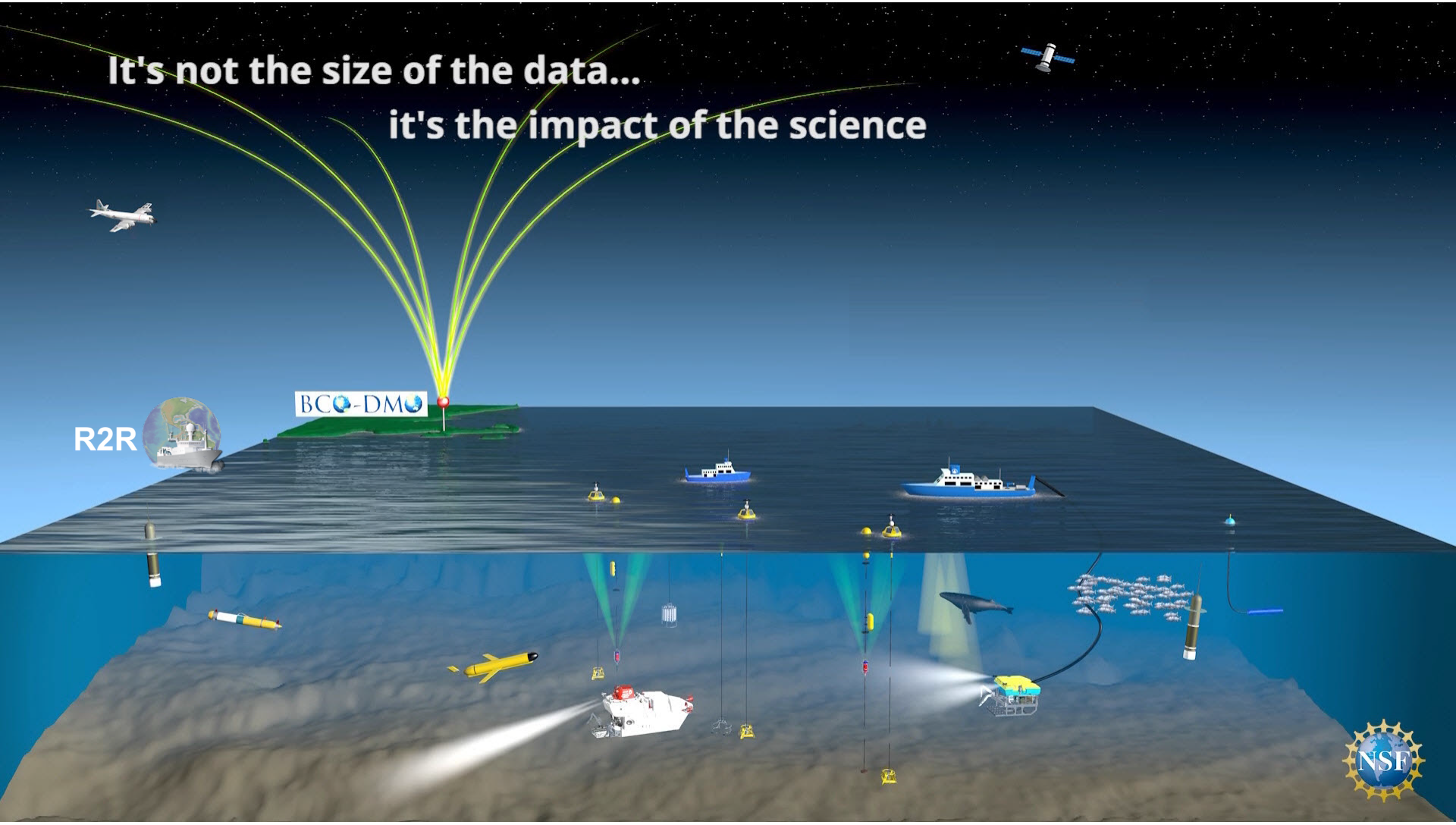
SEA STORY: PART 1

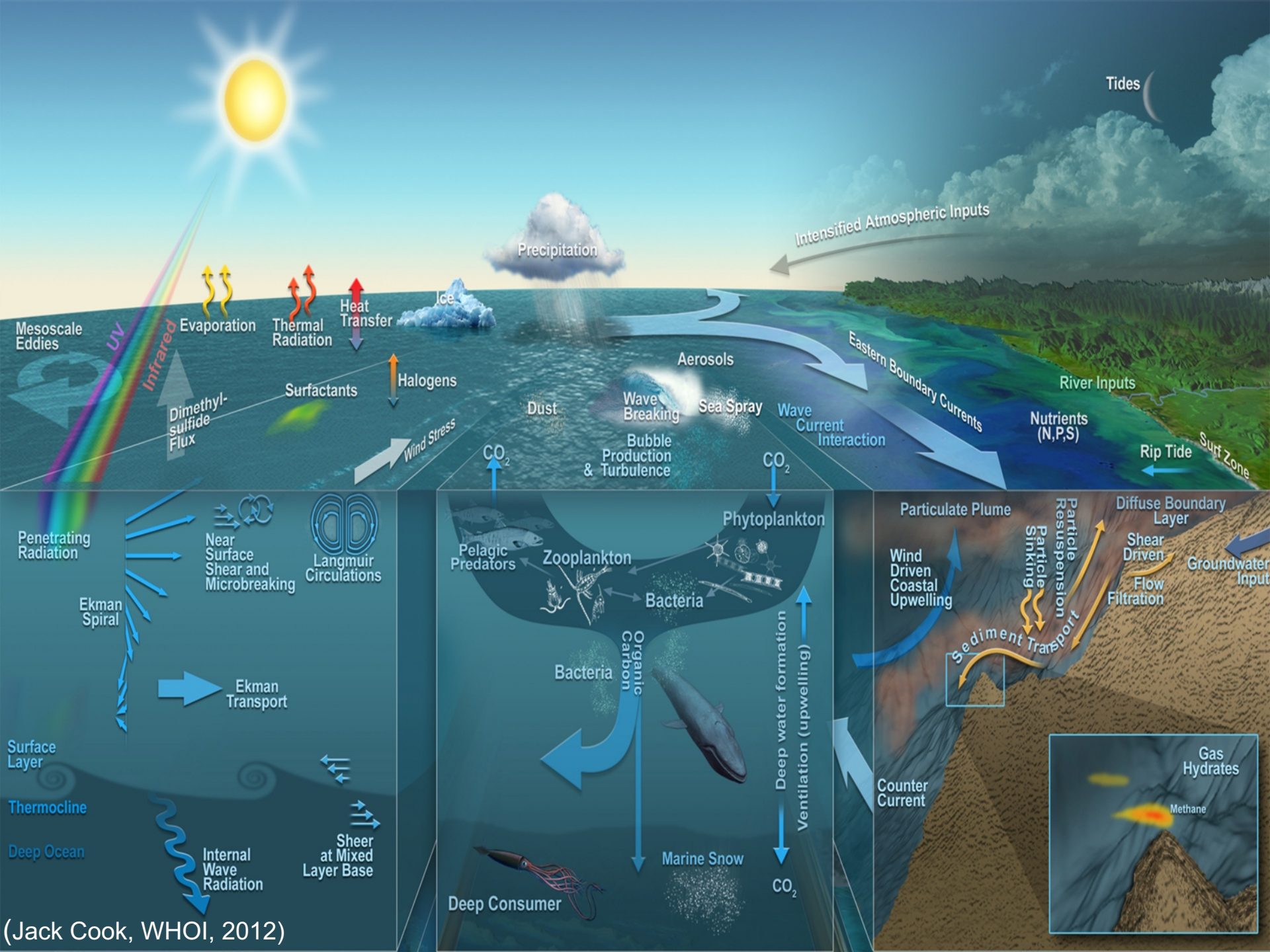


Google North Atlantic Fishery

North Atlantic Cheap
www.NexTag.com
Want Deals for **North Atlantic**? See NexTag Sellers' Lowest Price!

DISTRIBUTED DATA … GLOBAL SCIENCE

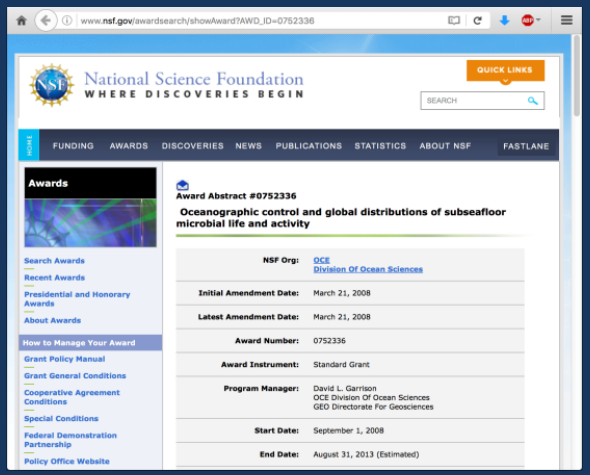




(Jack Cook, WHOI, 2012)

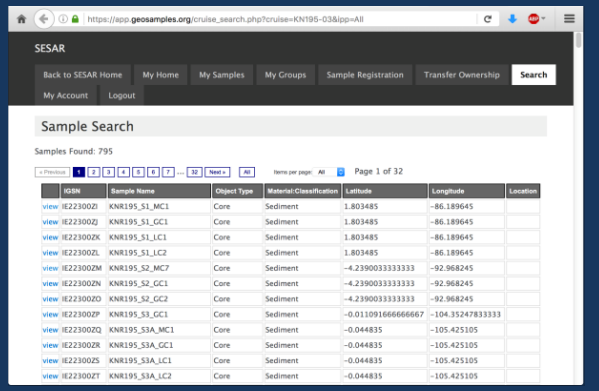
Distributed Related Resources

Funding Source

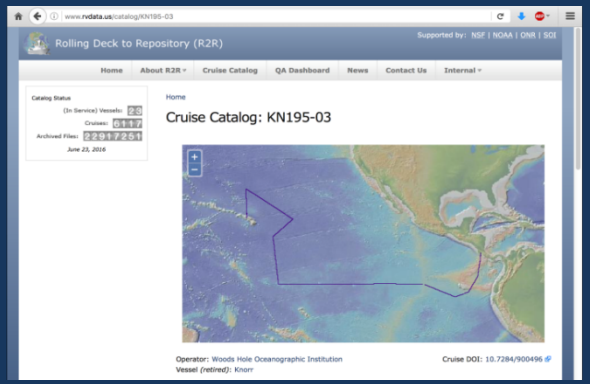


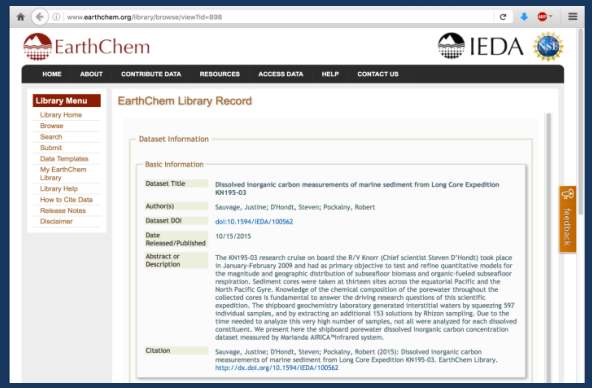
Data Sets

Physical Samples

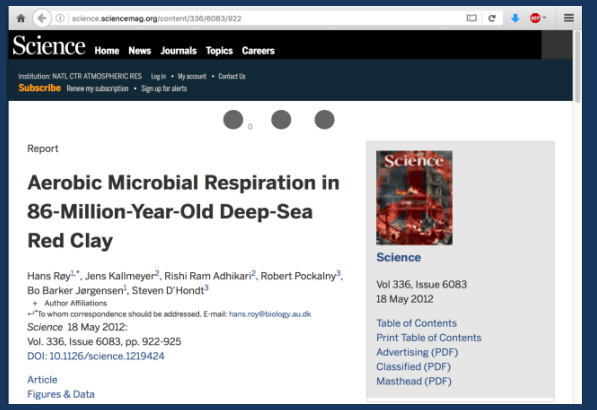


Field Expedition (e.g. cruise)





Publications

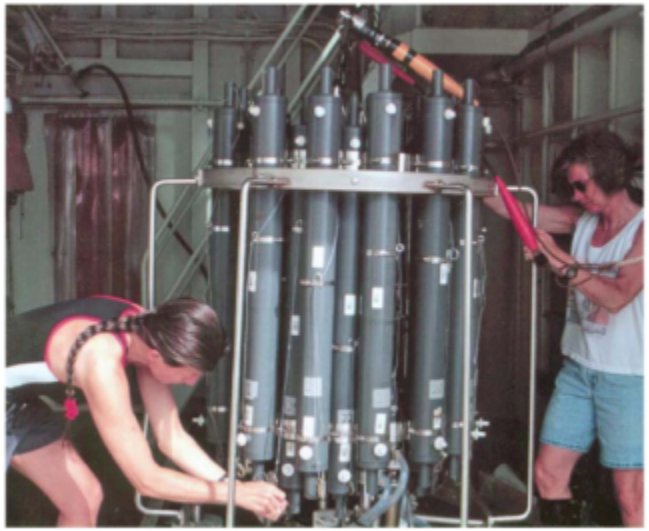


R2R, BCO-DMO, DATAONE

- 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)

Shared notes: http://bit.ly/S4S2017

PROGRESS SO FAR . . .

Metadata, plus

Semantic markup, plus

Persistent identifiers . . .

PROGRESS SO FAR . . .

Metadata, plus

Semantic markup, p

Persistent identifiers

Linking Related Resources via PIDs

Person @ORCID

Cruise @R2R

Award @NSF

Paper

Samples @SESAR

Dataset @ECL

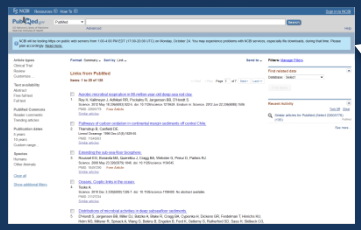
Dataset @BCO-DMO

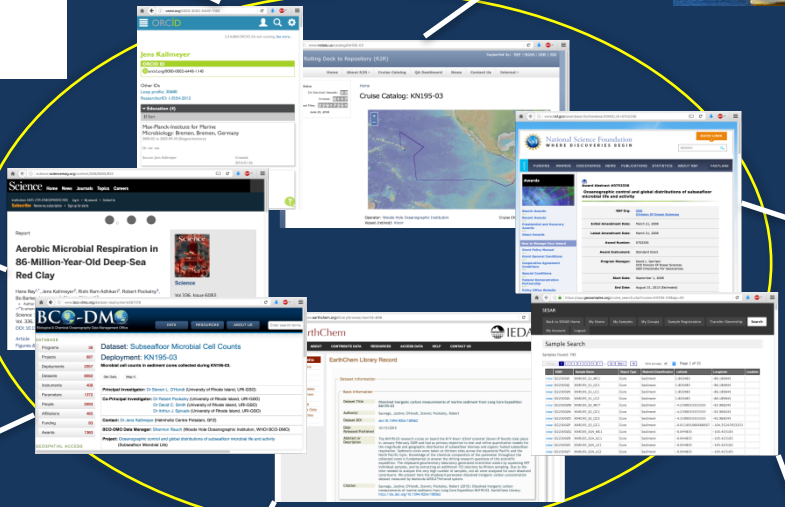
Linking Additional Resources

educational history

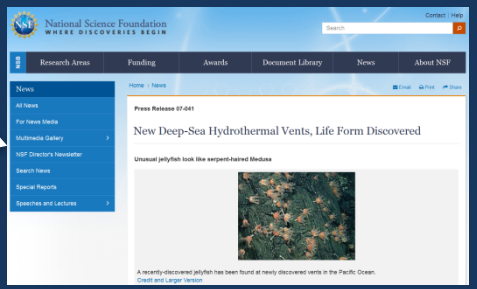
research vessel information

related articles





related research



instrumentation details

related analyses

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

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

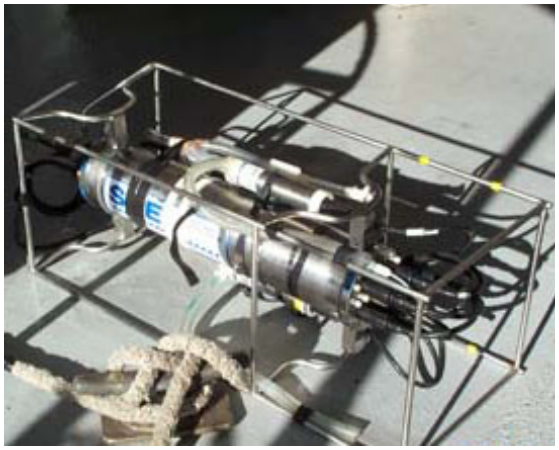
What other ways are people using semantics to meet information management challenges?

Provenance?

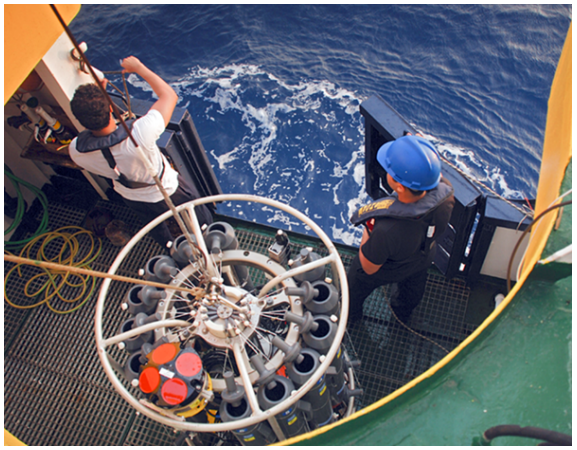
CONSIDERATION

Should we publish PIDs for controlled vocabulary terms?

e.g. classes of instruments



sameAs



Shared notes: http://bit.ly/S4S2017

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

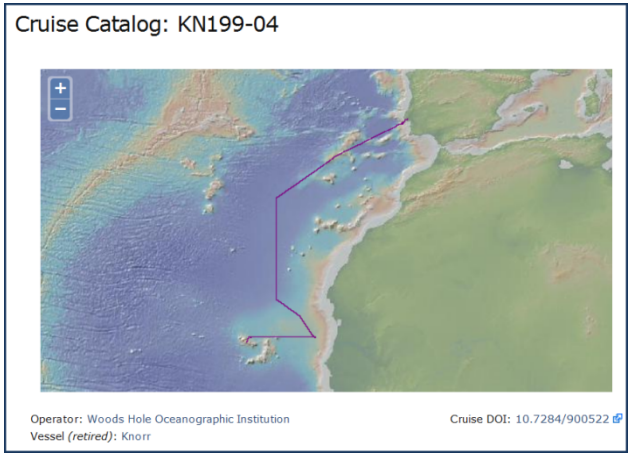
CONCERN

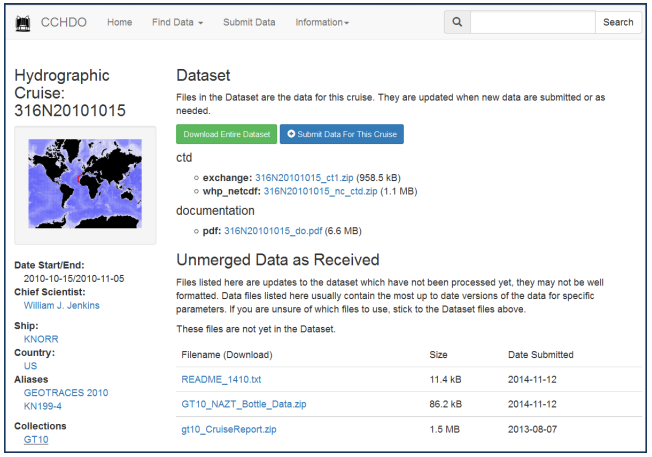
DOI assignment practices

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)

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





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?

Shared notes: http://bit.ly/S4S2017

CONCERN

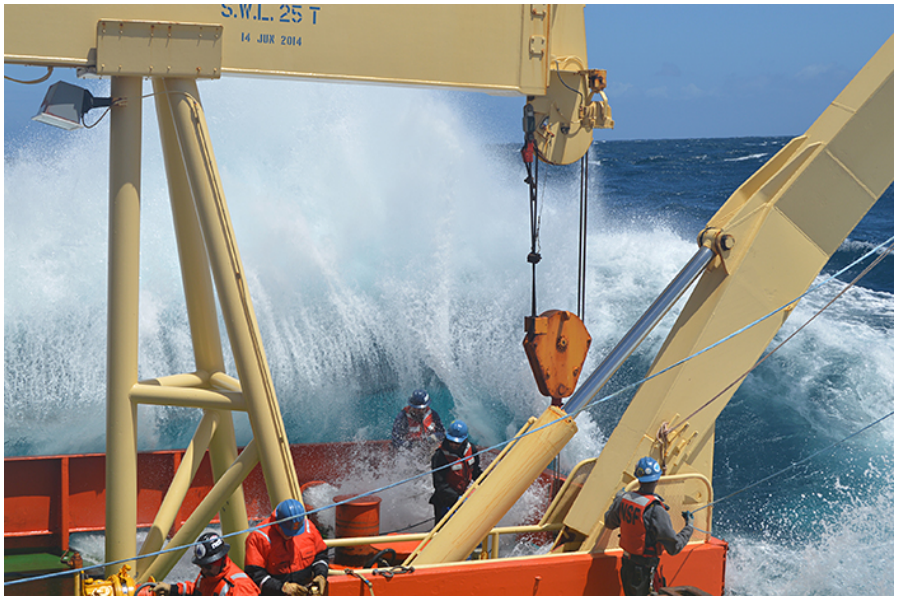
A DOI is a DOI … or are CrossRef and DataCite data set DOIs different ?

One of these things is not like the other …

but is there a lossless crosswalk between their metadata schemas?

Shared notes: http://bit.ly/S4S2017

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/