

Management of NSF Funded Ocean Research Data

CYNTHIA CHANDLER

BIOLOGICAL AND CHEMICAL OCEANOGRAPHY
DATA MANAGEMENT OFFICE

WOODS HOLE OCEANOGRAPHIC INSTITUTION

2016 US CLIVAR PSMIP Annual Meeting . Woods Hole, MA . 15 June 2016



CO-AUTHORS ON THIS WORK

Molly D. Allison, Nancy Copley, Stephen R. Gegg, David M. Glover,
Robert C. Groman, Danie Kinkade, Shannon Rauch,
Adam Shepherd, Peter H. Wiebe and Team BCO-DMO

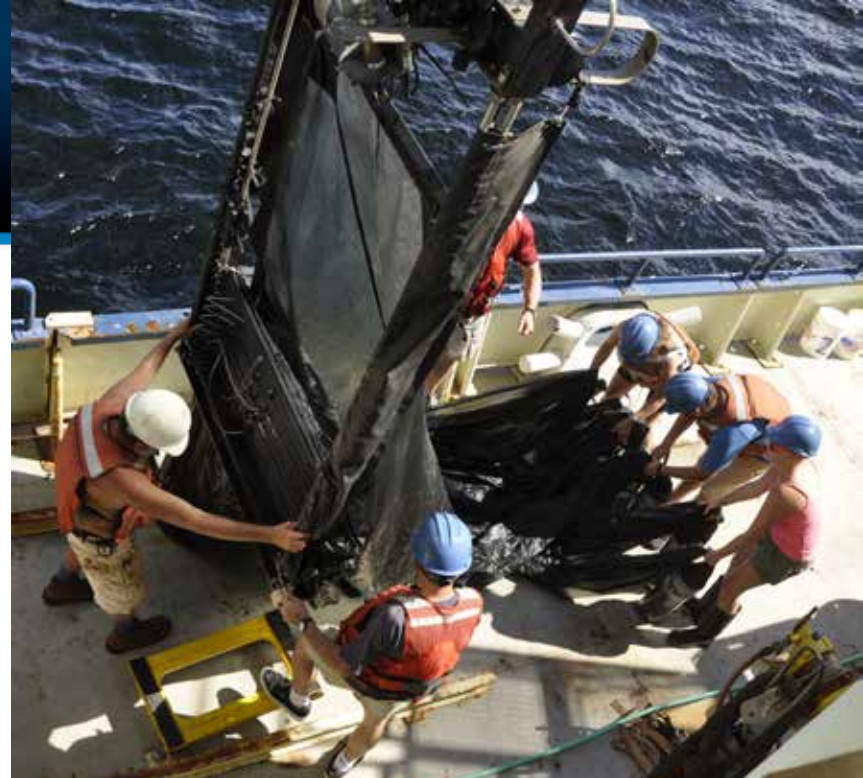
Biological and Chemical Oceanography Data Management Office
B C O - D M O



OUTLINE

Changes, Challenges & Recommendations

from my perspective . . .
as a manager of NSF funded
marine science data from
hypothesis-driven research





CHANGING RESEARCH PARADIGM

It is no longer enough to collect data by yourself, publish the paper and move on to the next research question.

We see greater expectations from funding agencies, researchers, and the extended community for open data access and **machine access**.

D. McGuinness, Fall AGU 2012, Community Science - The Next Frontier



THE DATA REVOLUTION

“[The data revolution] isn’t just about the volume of scientific data; rather, it reflects a fundamental change in the way science is conducted, who does it, who pays for it and who benefits from it. And most importantly, the rising capacity to share all [these] data – electronically, efficiently, across borders and disciplines – magnifies the impact.”

The Data Harvest: How sharing research data can yield knowledge, jobs and growth (RDA Europe, December 2014)

<https://europe.rd-alliance.org/>



NSF RESEARCH DATA MANAGEMENT

How does research data become resource data ?
Documentation is essential.

- BCO-DMO* is a domain-specific repository for NSF funded marine ecosystem data
- Data managers work with original PIs to add metadata (structured documentation)
- Standards compliant metadata to support access by machine clients

More documentation = better access to data

* BCO-DMO <http://bco-dmo.org>
Biological and Chemical Oceanography Data Management Office



WHAT DO RESEARCHERS NEED?

All types of researchers:

other science domains, educators, policy makers

- Access to data and metadata (documentation)
- Access to vast array of data types
- Cross-disciplinary research requires access to:
 - Discipline-specific metadata including quality control and provenance information
 - Published in machine-interpretable way
 - On-demand data visualization and integration
 - Data synthesis products (e.g. seasonal averages)

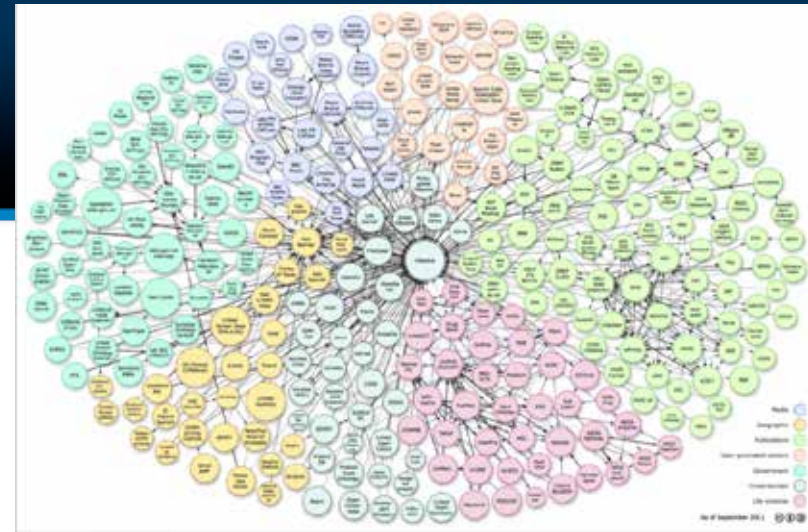
TRENDS

Linked Data Semantic Web Provenance

Hills, D. J., R. R. Downs, R. Duerr, J. C. Goldstein, M. A. Parsons, and H. K. Ramapriyan (2015), The importance of data set provenance for science, *Eos*, 96, doi:10.1029/2015EO040557. Published on 4 December 2015.

Quality Assessment, Quality Control

Documentation should capture what was done to ensure the best accuracy and precision from acquisition through final data product



<http://linkeddata.org>

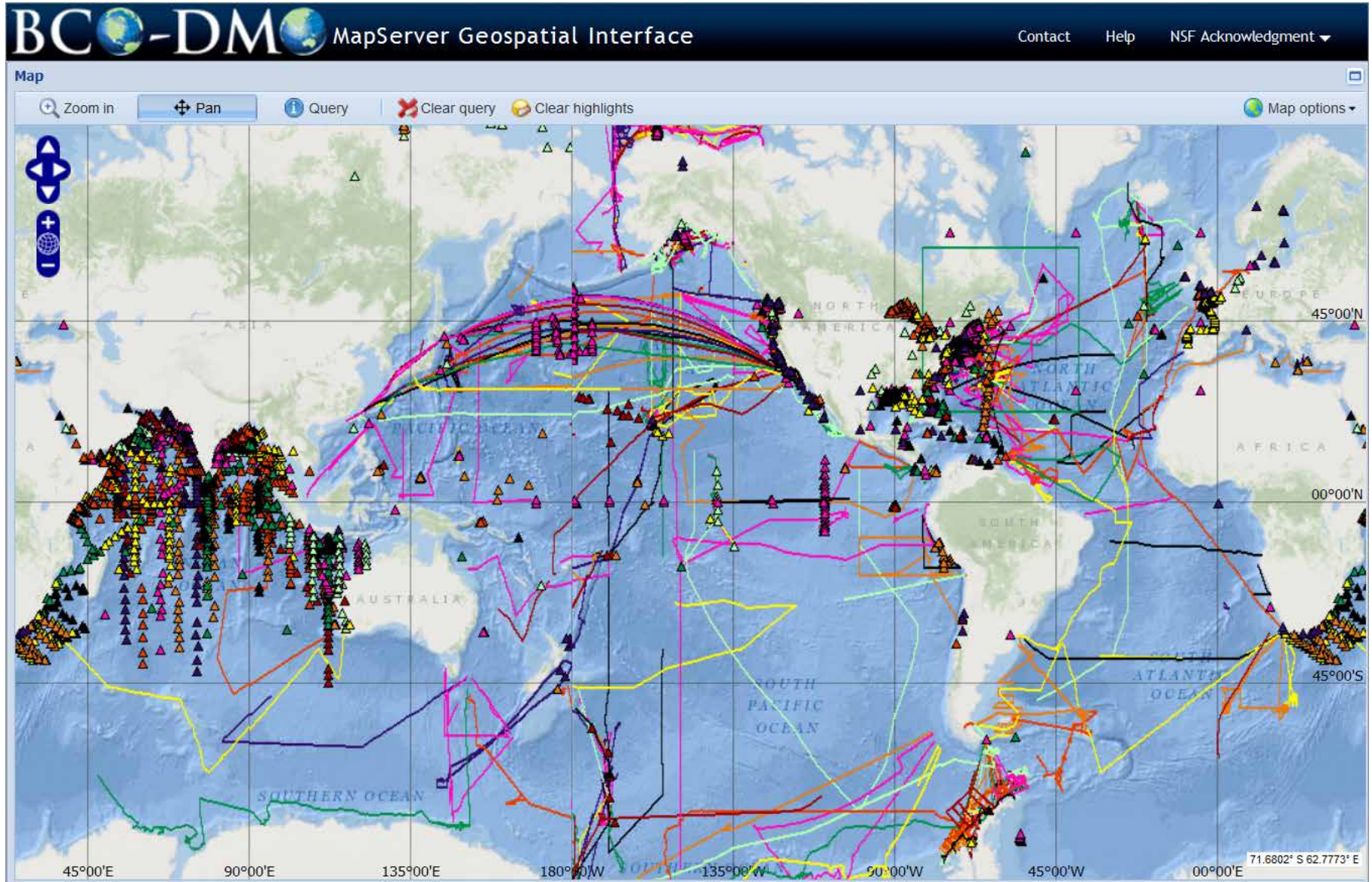


BCO-DMO: PART OF THE SOLUTION

BCO-DMO is one data office in the US, funded to work in partnership with marine research scientists to help improve access to research data

- NSF OCE and PLR since 2006
- 5 year research grant (2014-2019)
- no cost to OCE and PLR funded PIs
- data archived at NCEI-Silver Spring

BCO-DMO MANAGED DATA



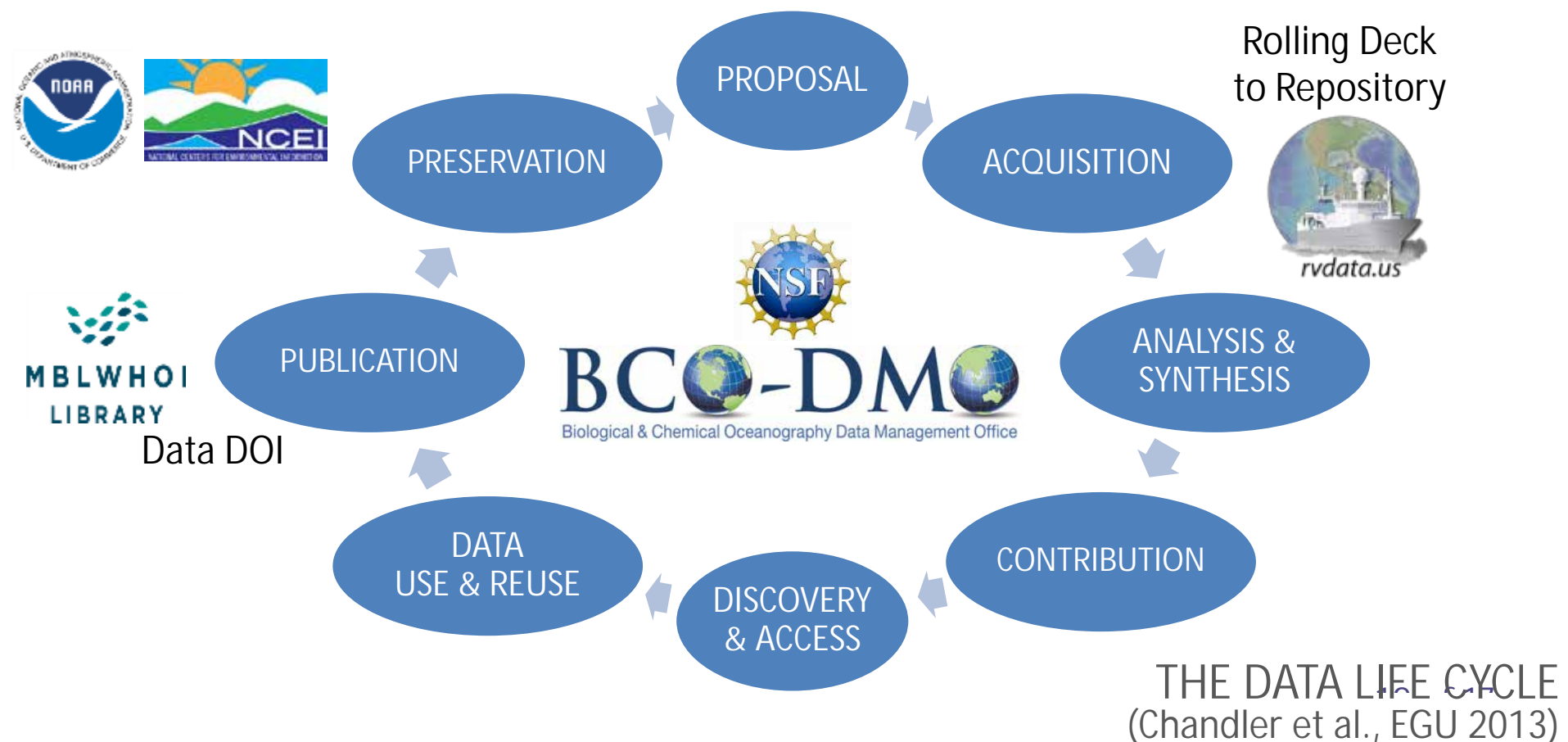


RECOMMENDATIONS

1. Intermediate data management facilities (sensor → analysis → archive)
 - Determine if/where they're needed
 - Create them
 - Commit to funding them
 - Use them to enable more/better research
 - Recognize the value of them
 - Work to improve them

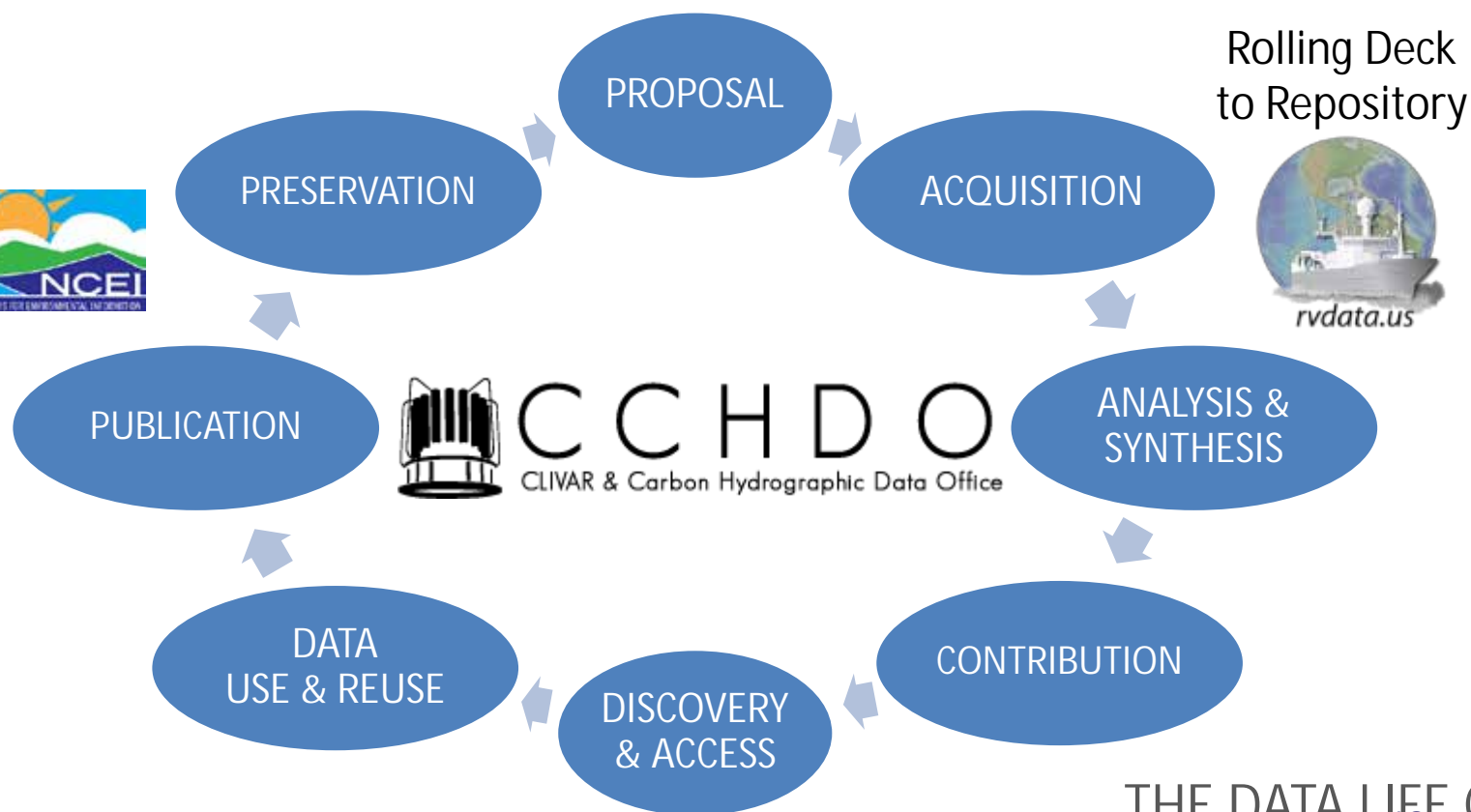
RECOMMENDATIONS

1. Intermediate data management facilities (sensor à lab à archive)



RECOMMENDATIONS

1. Intermediate data management facilities (sensor → lab → archive)



THE DATA LIFE CYCLE
(Chandler et al., EGU 2013)



RECOMMENDATIONS

- ## 2. Comprehensive community engagement
- JGOFS, GLOBEC, WOCE lessons learned
 - GEOTRACES (geotraces.org)
 - Data-Model Synergy Workshops
 - modified the sampling plan to support modelers
 - Inter-calibration activities
 - Pre-cruise planning workshops
 - modified the cruise track
 - Intermediate data products



RECOMMENDATIONS

3. metadata, metadata, metadata

- Data set documentation (provenance, QC)
- Research process
- Project and cruise

2. Methods

[5] The CLIVAR-CO₂ Repeat Hydrography program's A16N cruise was conducted aboard the NOAA research vessel Ron Brown which left Reykjavik, Iceland on 19 June 2003 and terminated in Natal, Brazil, on 12 August 2003. Details of the cruise track and ancillary data associated with the program can be found at http://cchdo.ucsd.edu/data_access?ExpoCode=33RO200306_01.

(Measures et al., 2008)

RECOMMENDATIONS

4. Share data early in the research process (see R1); improved quality, increased collaborative opportunities
5. Ensure persistent archive (NCEI) (see R1)
6. Formally publish the data (see R1, R5) assign a DOI (10.1575/1912/5258)
7. Cite data sources in papers (see R6)

THANK YOU

Comments . . . Questions?

Woods Hole, Massachusetts, USA



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

<http://bco-dmo.org>

cchandler@whoi.edu

[@cynDC42](#)

[@BCODMO](#)

ORCID: 0000-0003-2129-1647

EXTRA SLIDES

Extra, potentially relevant slides.

Slides after this point were not shown during the presentation.

CHALLENGE: How do we meet those needs?

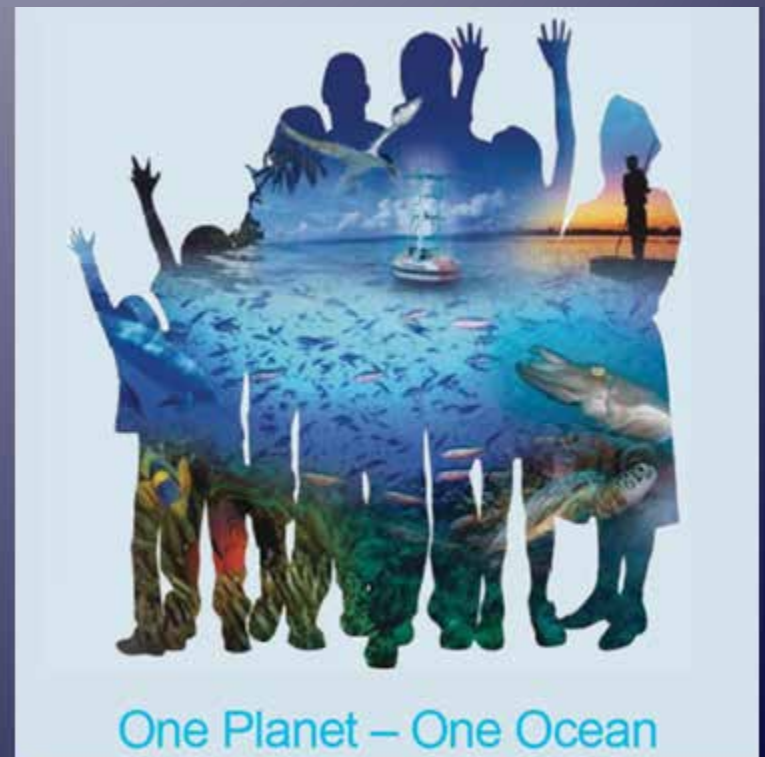
SOLUTION: Strategic partnerships to develop, share & adopt compatible strategies and solutions

§ Domain-specific

§ Cross-domain

§ Regional

§ International





SOLUTION: partnerships

COMMUNITIES OF PRACTICE

- US NSF EarthCube <http://earthcube.org/>
- ESIP <http://esipfed.org/>
- Research Data Alliance <https://rd-alliance.org/>
- ODIP <http://www.odip.org/>
- IODE <http://iode.org/>
- Future Earth <http://www.futureearth.info/>

BENEFIT FROM LONG-TERM COMMITMENT FROM AND
ACTIVE ENGAGEMENT OF AGENCY PROGRAM MANAGERS



- EarthCube
<http://earthcube.org/>

GOAL: to create integrated data management infrastructures across the US NSF geosciences by funding research efforts to complement, extend, enhance and connect existing infrastructure components.



ESIP

- Federation of *Earth Science Information Partners*
- <http://esipfed.org/>
an open networked community that brings together science, data and information technology practitioners



Research Data Alliance



- <https://rd-alliance.org/>
- supported by the European Commission, the National Science Foundation and other U.S. agencies, and the Australian Government; constructing the social and technical bridges that enable open sharing of data across technologies and between disciplines and nations with the ultimate goal of addressing the grand challenges of society.



Ocean Data Interoperability Platform

<http://www.odip.org/>

Goal: contribute to the removal of barriers hindering the effective sharing of data across scientific domains and international boundaries. ODIP welcomes all the major organizations engaged in ocean data management in EU, US, and Australia and is supported by the IOC/IODE.



IODE

- International Oceanographic Data & Information Exchange of IOC-UNESCO
- <http://www.iode.org/>
- International community of domain-experts
- A network of National Data Centers and Associate Data Units
- Infrastructure already in place to support marine science research community

Belmont Forum

- Belmont Forum: <http://www.bfe-inf.org/>
- established in 2009, brings together environmental and geoscience funding agencies from 15 nations and seeks to build a coalition of national resources to advance global environmental change research.



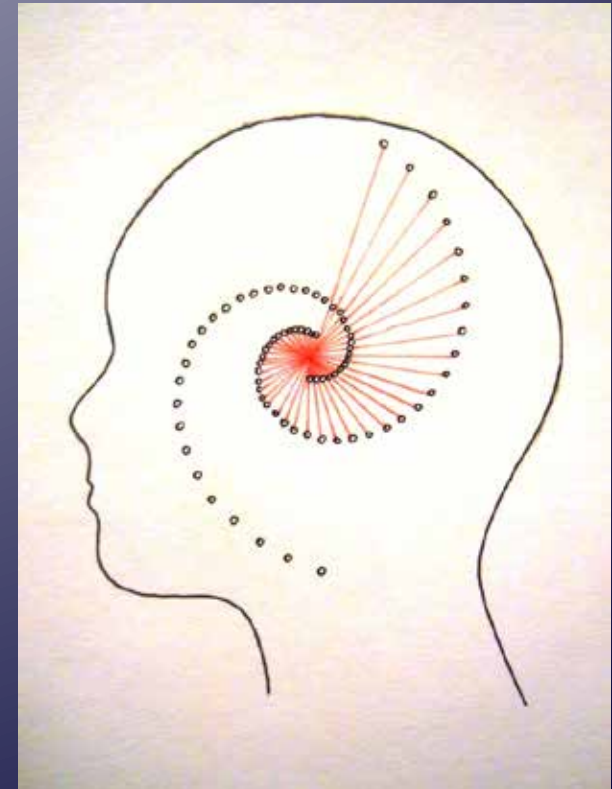
Future Earth



- <http://www.futureearth.info/>
- funding coordinated through the Belmont Forum, will be the platform through which many global change research programs will be coordinated, and the broad research themes, including the Earth Sciences, will require advanced information architectures to enable trans-disciplinary data-information-knowledge transfer

Challenge ~ Connectivity

- Goals & Strategies:
 - § linking content curated at distributed repositories
 - § improved interoperability (machine-to-machine)
- Technical solutions:
 - § metadata content standards
 - § controlled vocabularies
 - § Linked Data, Brokering
- Not just technical
 - § cultural conditions, behaviors
 - § research data lifecycle
 - § "proposal to preservation"






Strategy/Solutions: Connectivity

- Each 'fact' is explicitly declared and described by terms from (or linked to) community or global vocabularies
- Each term is identified by a globally unique Persistent Identifier (PID)
- Each PID resolves to a semantic representation of that term, with relationships to other terms
- Published as open, standards-compliant records

Challenge #2 ~ Legacy Metadata

- Human readable text metadata records from environmental, legacy data systems

Directory Data Display



HPLC_pigments

PI: Robert R. Bidigare
dataset: Pigments, HPLC method, sampled from bottle casts
cruise: TTN-045, Arabian Sea Process cruise #2
ship: R/V Thomas Thompson

[Methodology and Notes](#) revised 05/30/97

Parameters	Descriptions	Units
event	event number, from event log	
sta	station number, from event log	
sta_std	Arabian Sea standard station identifier	
cast	cast number, from event log	
bot	rosette bottle number	
depth_n	nominal sample depth	meters
chlide_a	Chlorophyllide a	nanogram/liter
chl_c3	Chlorophyll c3	nanogram/liter
chl_c	Chlorophyll c1 + chlorophyll c2 + Mg 3,8 divinyl pheoporpyrin a5	nanogram/liter
peridinin	Peridinin	nanogram/liter

PI: Robert R. Bidigare
dataset: Pigments, HPLC method, sampled from bottle casts
cruise: TTN-045, Arabian Sea Process cruise #2
ship: R/V Thomas Thompson

[Methodology and Notes](#) revised 05/30/97

chl_c3
chl_c
chlide_a

Chlorophyll c3
Chlorophyll c1 + chlorophyll c2 +
Mg 3,8 divinyl pheoporpyrin a5
Chlorophyll a plus chlorophyllide a



Strategy: Metadata for Smart Data

- Use the **Semantic Web** to connect (or link) distributed data repositories
 - Formal ontology modeling
 - Create ontology design patterns
 - Use controlled vocabulary terms
 - Publish content as Linked Data

C. Bizer, T. Heath and T. Berners-Lee. 2009. "Linked Data - The Story So Far", International Journal on Semantic Web and Information Systems, Vol. 5(3), Pages 1-22. [dx.doi.org/10.4018/jswis.2009081901](https://doi.org/10.4018/jswis.2009081901)

Solution: Metadata for Smart Data

- Publish as ISO 19139, W3C DCAT, schema.org Dataset extension
- Formal data publication with a DOI
- RDF with semantic markup including PROV



SHARED STANDARDS



Big Challenges & Big Opportunities

- Complex, large scale research questions
- Infrastructure (people, machines, systems) must be updated to support new research requirements
- Increased need for robust, discipline-specific, machine-actionable information (semantics)
- Standards-compliant metadata
- Documenting open-access data with PIDs (DOI)

Grand Challenge

- How to keep up with rapidly changing needs and expectations of
 - § Research community
 - § Funding agencies
 - § Educators
 - § Policy makers
 - § Other stakeholders
- ... given the usual limitations:
funds, time, skilled personnel
- ... and do it in a sustainable way?

CHALLENGE: access to a variety of data

SOLUTION: Semantic Web technologies

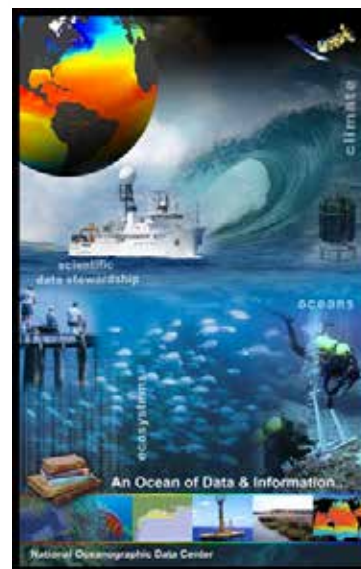
- Metadata standards w/ semantic markup
- Controlled vocabulary lists for efficient search
- Linked Data to connect related resources



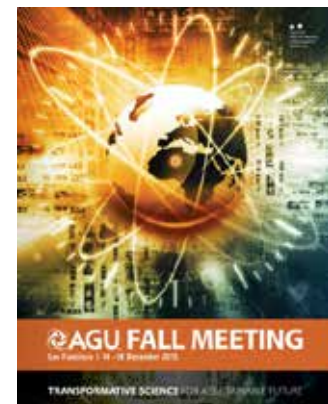
LTER



NATIONAL OCEANOGRAPHIC
DATA CENTER (NODC)
UNITED STATES DEPARTMENT OF COMMERCE



MBLWHOI
LIBRARY



NCBI

Ocean
CO₂
CDIAC



GULF OF MEXICO RESEARCH DATA



DATABASE

New Entry	
Programs	36
Projects	641
Deployments	2293
Platforms	464
Datasets	7847
Instruments	392
Parameters	1366
People	2007
Affiliations	469
Funding	78
Awards	1293

Projects registered at BCO-DMO

include links to other repositories:

Project-specific Web Sites:

research program descriptions

[R2R](#): original cruise data

[NCBI](#): sequence data



DATA & RELATED RESOURCES

- We need to be able to connect other resources besides the data ...
- **Scenario:** A publication by Gulf of Mexico researcher Samantha Joye says the data are at BCO-DMO, and you want to discover related data, meeting presentations, published papers, and the relevant funding sources, and ... More!

EXAMPLE: GEOLINK for GULF DATA



Rolling Deck to Repository (R2R)

Cruise Catalog



▶ 1. NSF funds a Gulf research cruise. 2. R2R serves the original underway data.

3. BCO-DMO manages and preserves the post cruise data, with links to NCBI.



▶ 4. meeting abstracts and 5. formal publications (data and peer-reviewed papers) complete the research data life cycle.



bco-dmo.org

If you know where to look, resources are discoverable at each repository, but what if we could connect them?

EXAMPLE: SAMANTHA JOYE

BC-DM
Biological & Chemical Oceanography Data Management Office

HOME DATA RESOURCES ABOUT US

DATABASE

- New Entry
- Programs 36
- Projects 639
- Deployments 2291
- Platforms 462
- Datasets 7839
- Instruments 391
- Parameters 1366
- People 2003**
- Affiliations 469
- Funding 78
- Awards 1291
- Contact Status 270
- Archive 170

Account: cchandler

Dr Samantha B. Joye

View Edit Delete Track contacts Add new person

Synonyms: Mandy
Affiliation: University of Georgia (UGA)
Address:
Department of Marine Sciences
Room 220 Marine Sciences Building
Athens, GA
USA 30602-3636
Email: mjoye@uga.edu
Phone: (706) 542-5893
Fax: (706) 542-5888

Activities reported for Dr Samantha B. Joye:

- ▶ Program coordination
- ▶ Project coordination
- ▼ Deployment coordination

More from GeoLink  

Deployment	Synonyms	Platform	Role
AT18-02		Atlantis	Chief Scientist
WS1010		F.G. Walton Smith	Chief Scientist

▶ Datasets

GEOLINK SEARCH RESULTS

More from GeoLink



R2R Cruises:

WS1010, AT18-02, AT26-13

DataONE:

22 GCE LTER datasets

BCO-DMO information:

2 Projects: DwHOS, Guymas Basin

Cruises: AT18-02, WS1010

Datasets: 4 available

NSF awards: OCE-1043225, OCE-1357360

bco-dmo.org @BCODMO



rvdata.us

Rolling Deck to Repository (R2R)

DataONE

BCO-DMO
Biological & Chemical Oceanography Data Management Office





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 19139, term URIs, links to other term URIs>

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