Re-envisioning data repositories for the 21st century

Diggs, S. and Kinkade D.

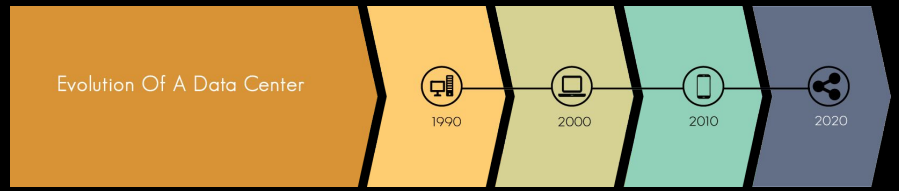
EGU General Assembly 2020
2020-05-07

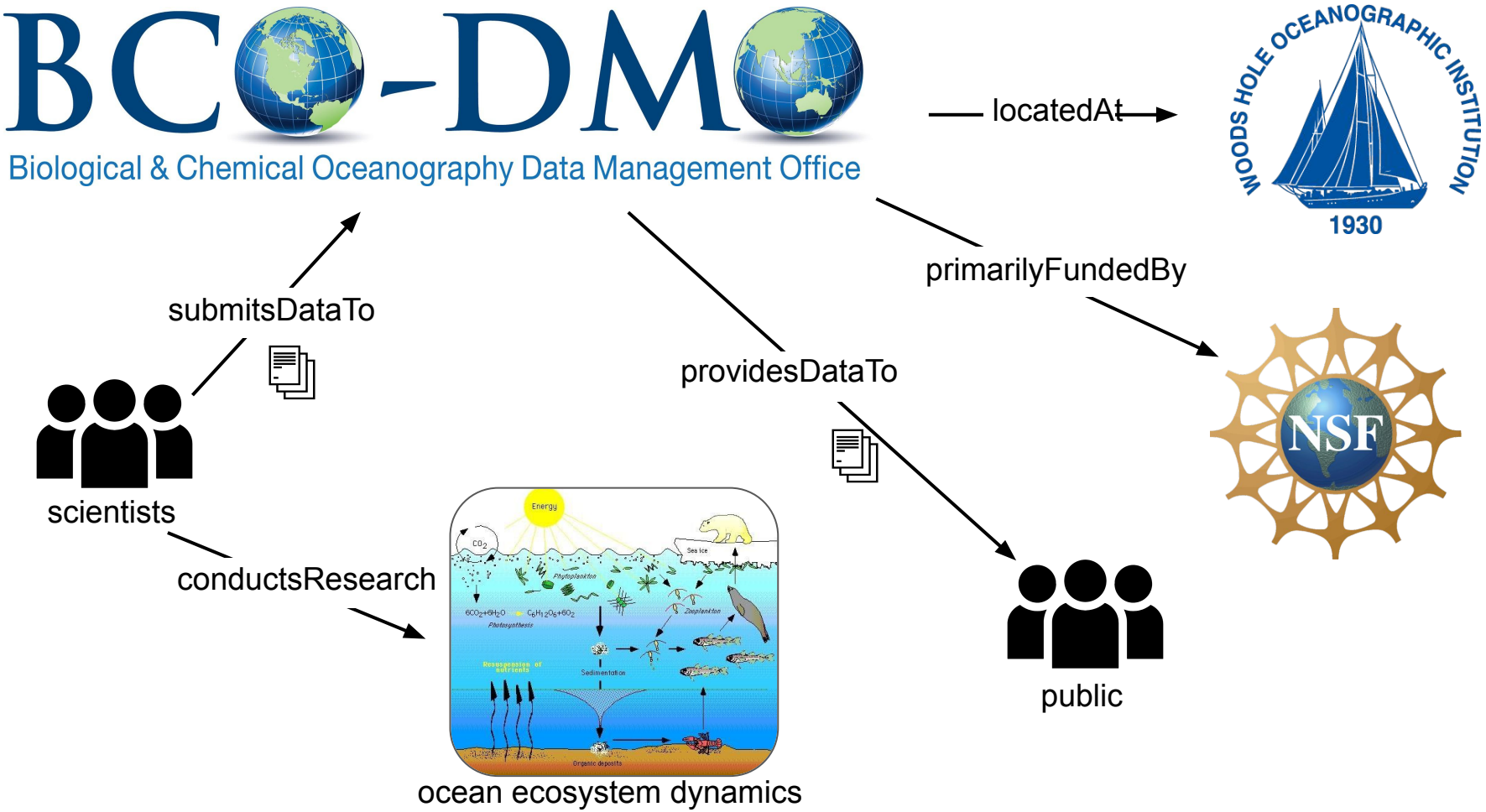
EGU2020-20826
https://doi.org/10.5194/egusphere-egu2020-20826

Credit: Octavio Aburto

FIFTY YEARS OF OCEAN PROFILE DATA

	1960	1970	1980	1990	2000	2010	2020
OCEAN SCI.		- GEOSECS	- SAVE (Atlantic) - CTD on Rosette - ADCP - TAO Array	- WOCE - LADCP - PALACE Floats	- CLIVAR - Argo Array - Pinniped CTD	- GO-SHIP	- Deep Argo - BIO-Argo
TECH / ORG	- NODC est. - Data by Mail - Inquiries by Phone - Acoustic Modem	- NOAA est. - DARPAnet - Service Argos - Minicomputers	- Personal Computers - Internet/Email - FTP - 9.6k Modem	- Civilian GPS - Email on Ships - WWW	- Internet on Ships - Iridium Satellites - Data Web Sites	- Personal Mobile Dev. - Data Web Sites - Cloud Storage - Social Media	
#of ocean profiles		3M	4.6M	6.4M	8.2M	12.14M	13.5M
disk cost/mb		$260	$192	$5.28	$0.01	---	
memory cost/mb		$7.4K	$6.4K	$98	$1	$0.01	




BCO-DMO
Biological & Chemical Oceanography Data Management Office
locatedAt
WOODS HOLE OCEANOGRAPHIC INSTITUTION
1930
primarilyFundedBy
NSF
submitsDataTo
providesDataTo
scientists
conductsResearch
Energy
Sea ice
Phytoplankton
Zooplankton
Sedimentation
Organic deposits
public
ocean ecosystem dynamics

BCO-DMO Data Types

Science Perspective

- Biological
- Chemical
- Biogeochemical
- Physical
- Geophysical

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

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

fromDiscipline

usedMeasurementTechnique

Highly Heterogeneous!

digitalDocument

storesValuesFor

DATASETS = >9000

Curation Perspective

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

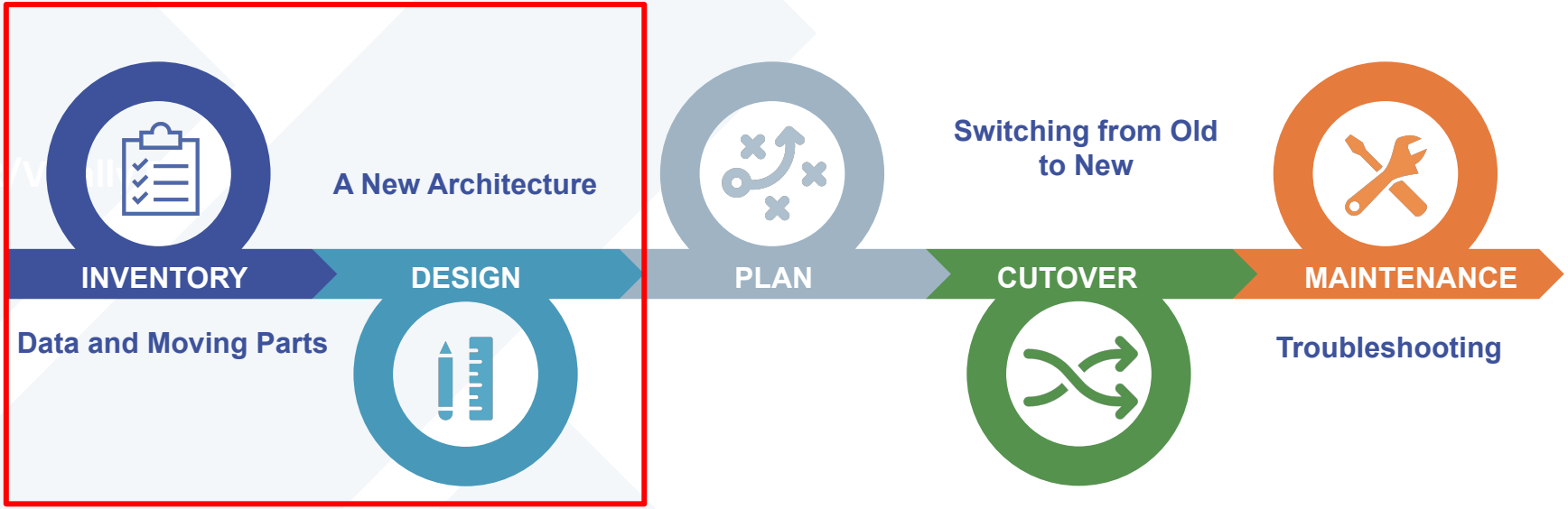
- Variable organization
- Varying metadata
- Local parameter terms
- Emerging data types
- Distributed complementary info

Many repository improvements are driven by the state of technology.

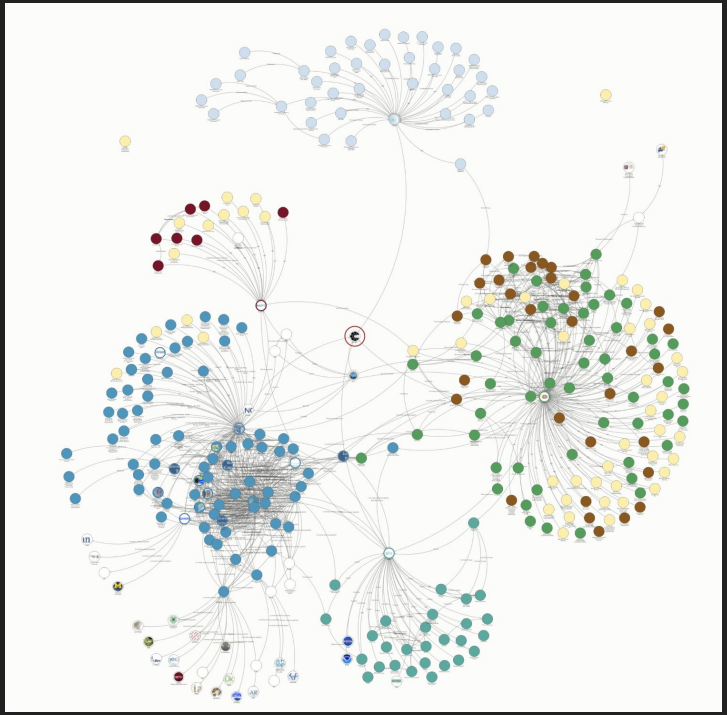
This is reactive.

How do we become proactive?

GENERALIZED REPOSITORY TRANSITION MODEL (IOOS)



Intentional evaluation of current technologies and best practices



BCO-DMO

Biological & Chemical Oceanography Data Management Office

Submission Software
Curation Software
Access Software
Custom APIs

currentlyDefiningModel

definesMetadataModel

DigitalObject Submission

DigitalObjectAccess

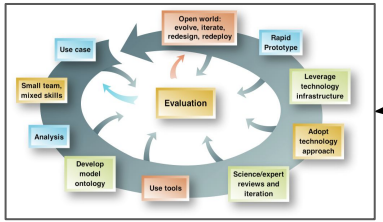
reducesAssumptionsInCode

DigitalObject

Provenance

describeWhatWeDoAndKnow

informedBy



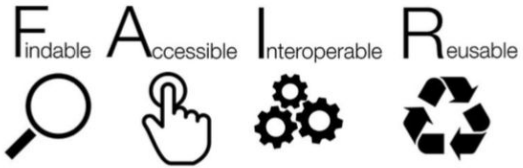
Best Practices for Publishing Linked Data

Data on the Web Best Practices

Schema.org Value

Why add additional information to websites?

- Improve *findability*
  - Google and other indexers
- Facilitate data *access*
  - Programmatic support for resource access
- Enhance *interoperability*
  - Well described resources
- Promote *reuse*
  - Simplified programmatic access

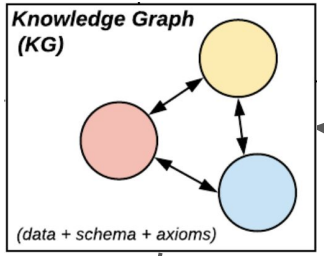


Result: new data model reflects the current state of original and adjacent domains.

BCO-DMO

Biological & Chemical Oceanography Data Management Office

lod.bco-dmo.org



usesModel

- Dublin Core
- Dataset Catalog (DCAT)
- PROV-O
- Schema.org
- Bibliographic Ontology
- GeoSPARQL
- NERC Parameter Usage Vocabulary
  - Environment Ontology
  - CHEBI
  - *and more...*
- *NanoPublications*
- *Ocean Data Ontology (BCO-DMO)*

followsPattern

supportsQueriesFor

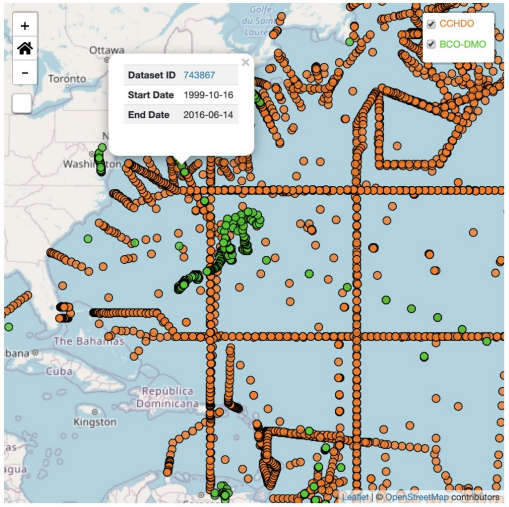
- Dataset & Data Type
- Funding Source & Award
- Geospatial & Temporal Coverages
- Instrumentation
- Observed Properties
- People & Roles
- Persistent Identifiers (DOIs, ORCiD, etc.)
- Related Publications (by PID too)
- Research Topic
- Supplemental Documents

constructedWith

RDF

- Science-on-schema.org (ESIP)
  - Dataset
  - Repository Services
- Data Type Registry (RDA)
  - Ocean Proteomics
  - *more coming*
- *PID Kernel Information Profile (RDA)*

Coordination of technologies and practices lay a solid foundation for desirable outcomes



1950 2020

CCHDO Bottle CTD Data

Show 10 entries Search:

Expocode	Start Date	End Date
18HU20130507	2013-05-07	2013-05-28
18MF20120601	2012-06-01	2012-06-17
33AT20120419	2012-04-19	2012-05-15
33AT20120324	2012-03-24	2012-04-17
316N20111106	2011-11-06	2011-12-11
18HU20110506	2011-05-06	2011-05-29
18HU20100513	2010-05-13	2010-05-30
06M220090714	2009-07-15	2009-08-12
18HU20090517	2009-05-17	2009-06-01

BCO-DMO CTD Data

Show 10 entries Search:

Dataset Id	Start Date	End Date
3358		
781545	2019-07-25	2019-07-25
774859	2019-05-20	2019-05-23
774958	2019-05-20	2019-05-23
757784	2018-05-07	2018-05-29
757722	2017-10-14	2017-10-27
753679	2017-03-30	2017-04-04
753624	2016-09-07	2016-09-09
614744	2014-06-19	2015-08-28

The Journey Continues…

@scdiggs
@danie_b_k