Navigating a Sea of Big Data

BCO-DMO
Biological & Chemical Oceanography Data Management Office

bco-dmo.org

IN11C-3622
AGU Fall 2014

Danie Kinkade, Cynthia Chandler, Robert Groman, Adam Shepherd, Molly Allison, Shannon Rauch, Peter Wiebe, David Glover • Biological and Chemical Oceanography Data Management Office, Woods Hole Oceanographic Institution (WHOI), Woods Hole, MA USA

Abstract

Oceanographic research is evolving rapidly. New technologies, strategies, and related infrastructures have catalyzed a change in the nature of oceanographic data. Heterogeneous and complex data types can be produced and transferred at great speeds. This shift in volume, variety, and velocity of data produced has led to increased challenges in managing these Big Data. In addition, distributed research communities have greater needs for data quality control, discovery and public accessibility, and seamless integration for interdisciplinary study. Organizations charged with curating oceanographic data must also evolve to meet these needs and challenges, by employing new technologies and strategies.

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) was created in 2006, to fulfill the data management needs of investigators funded by the NSF Ocean Sciences Biological and Chemical Sections and Polar Programs Antarctic Organisms and Ecosystems Program. Since its inception, the Office has had to modify internal systems and operations to address Big Data challenges to meet the needs of the ever-evolving oceanographic research community. Some enhancements include automated procedures replacing labor-intensive manual tasks, adoption of metadata standards facilitating machine client access, a geospatial interface and the use of Semantic Web technologies to increase data discovery and interoperability.

BIG DATA PRESSURES

drive facility modifications and adaptations

The evolution of the Big Data Vs exerts pressure on intermediate data facilities. When existing systems and tools no longer support the research community's current needs, a facility is dealing with Big Data!

Before 2006

Data management was conducted by project-specific, locally-managed data offices. Content veracity was determined by hand.

2006 V V

BCO-DMO is created to meet the data management and sharing needs of thematic, interdisciplinary researchers funded by the **National Science Foundation's Chemical and Biological Oceanography Divisions** and the **Division of Polar Programs.**

2007 V V

BCO-DMO adopts standards-based data management using a relational database **(mySQL)** to combine existing project metadata and data into a thematic repository.

2007 V V V

Using **OGC** services BCO-DMO works with **Second Creek** to develop data visualization via a standards compliant, interactive geospatial interface to discover and display an ever increasing variety of data types.

2008 V V V

BCO-DMO develops a controlled vocabulary with the **Rolling Deck to Repository project** to achieve standardization and interoperability with other community information resources.

2009 V V V

BCO-DMO employs semantics collaborating with **RPI** to construct an ontology, allowing users to discover a variety of data, as well as assisting in quality control of content.

2010 V V V

BCO-DMO automates ingest of NSF funding award information, allowing data managers to rapidly add crucial metadata related to projects funded by the **National Science Foundation** into the database.

2010 V V V

BCO-DMO partners with the **MBL/WHOI library** for data publication. DOI creation facilitates discovery of varying data types, and recognition for the investigator.

2010 V V V

BCO-DMO develops automated submission procedures to submit data holdings to the **National Oceanographic Data Center, NODC,** to relieve manual submissions of large data volumes.

2013 V V V V

BCO-DMO migrates its metadata database from ColdFusion to **Drupal**, an open source content management system, providing increased flexibility to meet emerging web-based challenges.

2014 V V V V

BCO-DMO adopts the **ISO 19115-2** metadata standard to increase interoperability of metadata information among like repositories, and content quality.

2014
434 PROJECTS

2014
7193 DATASETS ONLINE
comprising in situ, satellite, video, experimental and model data

2006
19 PROJECTS

2006
2574 DATASETS ONLINE
Comprising in situ cruise data

RESPONSES

BIG DATA DRIVERS

Before 2006

Data and information were small in size and kept with the investigator, shared in an ad hoc manner with collaborators.

2006

NSF encourages collaboration of individual oceanographic research facilities to serve thematic research community needs.

2007

Collection and management of multiple project metadata becomes unscalable using human-generated, and readable flat files.

2008

Variety of data parameters being collected and submitted becomes too large to manage in an ad hoc manner.

2009

New oceanographic research community needs require networked repositories for discovery of interdisciplinary data. Higher need for content quality.

2009

Semantic Web technologies drive the need for more structured content management.

V Data Veracity
Increased variety and volume of data pose a problem for validating the quality and veracity of data and information. Using semantics and linking to trusted, authoritative partners can facilitate quality control of content.

V Data Velocity
Increased speed in which data arrive for processing taxes the data management system. Expectation for community data sharing requires rapid data processing, serving, and visualization.

V Data Variety
Increases in the variety of data sources as well as types of data found in each dataset, necessitate new strategies for management, and for discovery and access.

V Data Volume
Some varieties of data have become too large to manage with traditional technologies and strategies. Additionally, the total volume of data holdings becomes unwieldy for the user to discover and access data of interest.

Acknowledgements

BCO-DMO is funded by NSF grant OCE-1435578. We acknowledge all our partners, research collaborators, and the research investigators and their associates who have shared their data through BCO-DMO.