EXECUTIVE SUMMARY:
EARTHCUBE END-USER DOMAIN WORKSHOP REPORT

Submitted: 13 November 2013

Editors: Planning Committee Members

Workshop Title:
Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community

Woods Hole Oceanographic Institution
Woods Hole, MA
7-8 October 2013

PLANNING COMMITTEE

Danie Kinkade, (Chair) Woods Hole Oceanographic Institution (WHOI)
Cynthia Chandler (WHOI)
David Glover (WHOI)
Robert Groman (WHOI)
David Kline (Scripps Institution of Oceanography, University of California, San Diego)
Jasmine Nahorniak (Oregon State University)
Todd O'Brien (NOAA, National Marine Fisheries Service)
Mary Jane Perry (School of Marine Sciences, University of Maine)
James Pierson (University of Maryland Center for Environmental Science)
Peter Wiebe (WHOI)

INTRODUCTION

An EarthCube Water Column Domain End-User Workshop hosted by the Biological and Chemical Oceanographic Data Management Office (BCO-DMO) was held October 7-8, 2013, at Woods Hole Oceanographic Institution. This executive summary synthesizes the workshop discussions, while further details such as use case descriptions and results of plenary and breakout group discussions are provided in the full workshop report.

The goal of the workshop was to articulate cyberinfrastructure needs of the ocean ecosystem dynamics community with particular focus on the challenges presented by multi-disciplinary marine ecosystem research. The ocean ecosystem dynamics domain encompasses a broad array of disciplines that often requires investigations in four dimensions, and seeks to increase understanding of the interplay between biological, chemical, and physical processes in the ocean. It is fundamentally an interdisciplinary domain by nature, producing highly diverse data types that pose unique challenges for management, integration, and analysis. The ability to discover, access, and synthesize high quality data from various disciplines is crucial to ocean ecosystem sciences.

The workshop brought together 50 participants (43 on-site and 7 remote) to explore and document the community's cyberinfrastructure needs from the domain users' viewpoint. The participants included 22 established, 16 mid-career, and 12 early-career or postdoctoral researchers. Individuals self-identified into one or more disciplines: oceanographers (68% of total), data and information managers (42%), cyberinfrastructure researchers (22%), education and social science specialists (14%), and modelers (8%). It should be noted that the timing of the lack in Federal appropriations and subsequent government shutdown prevented several registrants from participating in this workshop.

SCIENCE ISSUES AND CHALLENGES

1. Important science drivers and challenges: Participants identified several high-priority science questions that will drive interdisciplinary research efforts during the next 5-15 years.

- How will ocean acidification, warming, and hypoxia affect marine ecosystems? How do these phenomena affect the organisms? What species will be most impacted? What will be the impacts on ecosystem structure and dynamics?
- As ocean circulation changes in response to warming and changing salinity, how will marine organisms respond to resultant changes in their environment? What will be the impacts on productivity, species distributions, and carbon flux in the ocean? How will such ecosystem changes affect human populations?
- How will species distributions, life histories, and species interactions in polar oceans be impacted by the changes in ice cover? How will these polar changes impact other regions including changes in weather patterns? What will be the bottom up effects on marine coastal and open ocean ecosystems?
- How do different physical, chemical, and biological processes come together to create more complex emergent properties in ocean ecosystems? What are the feedback loops among these processes and how do they give rise to biological, chemical, and genetic diversity in the oceans?
- How are anthropogenic impacts other than climate change, such as eutrophication, overfishing, and coastal development influencing marine ecosystems? How does this influence our approach to ecosystem based management and conservation?

2. Current challenges to high-impact, interdisciplinary science:

A. Participants recognized three different categories of barriers preventing them from conducting transformative science: cultural, institutional, and technical.

Cultural barriers highlighted the reluctance to share data, a lack of confidence that data will be cited, different sharing expectations across domains, and the inability to access dark data (defined below in section B). Also mentioned were differences between domain vocabularies and an inability to translate science to the public.

Institutional barriers include the existence of organizational stove pipes that hinder cross-discipline collaboration, and a lack of the following: inter/intra-agency collaboration on data management, incentive and support for international and interdisciplinary collaborations and proposal reviews, recognition by tenure committees when submitting data for sharing, support for existing routine monitoring and observations, and insufficient funding for cutting edge science.

Technical barriers focused on aspects of data, such as lack of a single source or common method for discovery and access, and a lack of quality control assessments. Other data related barriers included difficulty with large data sets, combining heterogeneous data from different types of data streams, and difficulty dealing with evolving data types (such as 'Omics' data). Additional technical barriers were a lack of analysis code/tools for automation and visualization, and a lack of effective technological solutions to share data, etc.

B. Additional, more detailed challenges faced by participants:

Databases	Difficult to keep on top of all current cyberinfrastructure efforts (DataONE, EarthCube, Data Conservancy); lack of interoperability and diversity of data structures among data centers; lack of a common database capable of storing, searching, and serving terabytes of images and videos; submission of the same data to multiple repositories; difficulty knowing where to submit data.
Access	Lacking clearinghouses for (a) software/tools/code (b) databases (c) models, and (d) vocabularies/ontologies; difficult to find, access, and work with data outside of your area of expertise; difficult to access data from international databases; difficult if not impossible to access data from publications.
Dark data	No access to dark data (data on local computers or in file cabinets that are not in any public repository and not discoverable); challenges include lack of metadata, changing personnel, and lack of time and funding.
Data types	Some data are difficult to put into repositories (due to size, structure, format).
Metadata	Lack of rich, standardized metadata.
Quality	No standardized way to assess data quality
Tools	Lacking tools for automation, integration of a variety of data, and visualization in 4-D; a challenge to combine different data types and to deal with new and evolving data types; difficult to analyze heterogeneous data with gaps; a lack of systems for documenting provenance.
Hardware	Increasing need for data storage, processing capabilities, and bandwidth.
Education	Lack of data and computational literacy; need for data literacy courses; lack of online training tools for discovering, accessing, and using data.

Effort	Substantial time and cost in the effort of making data available. Reviewers are often unsympathetic to the high cost of data management activities.

TECHNICAL INFORMATION/ISSUES/CHALLENGES

Identified below are several critically needed tools, repositories, and infrastructure needed for pursuing key science questions:

1. Tools:

- Tools to enhance data discovery and searches.
- Visualization tools for interactive data analysis.
- Tools to track and control data versions.
- Tools to foster data quality assurance and quality control.
- A community-level interface and facility to share tools and programming code.
- Technology to assimilate metadata produced by smart sensors.
- A tool to translate from different format types to a standard format.
- Automatic incorporation of new data into databases/repositories
- Automatic retrieval of data for use in applications (e.g., forecasting)

2. Repositories and Databases:

- New data repositories (or expansion of existing facilities) are needed for emerging data streams that currently are not supported by a repository (e.g., metabolomics, citizen science data)
- A system for handling massive amounts of data including images and video.
- Production of a wider range of more sophisticated data products and derived calculations.
- Ensuring repositories function as, or work with archive facilities for long-term preservation of data.

3. Global Infrastructure:

- Guidance on where to submit data including restrictions and guidance for repository use.
- A centralized forum providing information about models, scripts, software, and documentation.
- Undergraduate and graduate-level curricula for training the next-generation of scientists to be able to find, submit, and work with data.
- Educate programmers to understand science.
- Standard interdisciplinary metadata format.
- Cross-domain ontologies of measurable phenomena and instrument types.
- Further development of crowd-sourcing funding and technology

COMMUNITY NEXT STEPS

Below is a list of tangible items and actions by EarthCube that would facilitate the community achieving its transformative science goals:

Short-term Next Steps (1-3 years):

- Catalog of different data repositories and tools.
- Catalog of existing and dark tools (for discovering, analyzing, and visualizing data).
- A tool that captures the output from multiple Earth System models for a geographical position for comparison to field-collected data.
- Better search tools for data discovery (such as faceted searches).
- Help desk to provide investigators with information on repositories where data should be submitted.
- Provide incentives to preserve all data and make accessible.
- Data and computational literacy curriculum including tutorials for undergraduate and graduates.
- Synthesis of outcomes from the domain-specific EarthCube workshops.

Long-term Next Steps (>3 years):

- Centralized access to earth system and ocean model output data with Google style searchability.
- Making databases and repositories interoperable.
- Plan to identify and liberate dark data (resources, funding, and expertise).
- Provide funds to enable Use Cases put forward in the workshop to be implemented.