EXECUTIVE SUMMARY: EARTHCUBE WORKSHOP RESULTS

Earth Cube Workshop Title: *Deep Seafloor Processes and Dynamics*

V.L. Ferrini (LDEO) and K. Rogers (CIW, RPI)

June 5-7, 2013

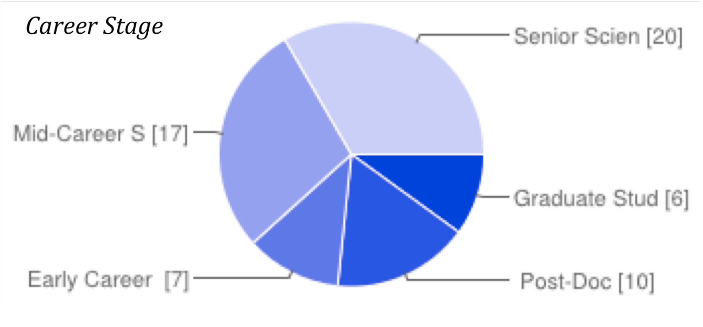
Introduction (field(s)/area(s) of interest and purpose, number of participants):

At the interface of Earth's interior and its external/surface environment lies the deep seafloor environment. The seafloor serves as the primary conduit for mass and heat transfer between sub-seabed and ocean systems, which operate on vastly different time and mass scales. Dynamics at this interface drive global (bio)geochemical elemental cycles, control global ocean chemistry, shape the surface atmospheric and climate system, and define the Earth's surface via tectonic processes. The seafloor-ocean interface also hosts some of the most diverse and extreme ecosystems in the biosphere, including deep-sea hydrothermal vents, cold seeps, mid-ocean ridges, deep-water coral ecosystems, ridge flanks, and plate margins, to name only a few. Research in deep seafloor processes spans a variety of disciplines as well - petrology, geology, geophysics, hydrogeology, aqueous geochemistry, micro- and macro-biology, ecology, and evolutionary biology - and the transformational science in deep seafloor systems occurs at the interface of these disciplines. True interdisciplinary research in deep seafloor dynamics requires mining and integration of large datasets from disparate disciplines and data integration and management are key components to the future success of interdisciplinary research in this field.

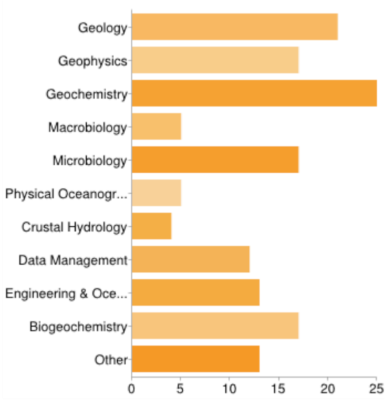
Scientists working in the deep seafloor environment comprise a model interdisciplinary end-user group that will benefit from the NSF EarthCube initiative. Integration of datasets generated by the deep seafloor research community could serve as a framework for analogous systems where integration of spatial and temporal cross-disciplinary data is crucial to the continued success of research efforts. The EarthCube End-User Domain Workshop for Deep Seafloor Processes and Dynamics targeted the major stakeholders in the deep seafloor research community and cyberinfrastructure specialists to chart the data integration and management needs into the EarthCube domain. As part of previous efforts to increase the participation of early career scientists in deep seafloor research, applications from graduate students, post docs and assistant professors and other early career scientists are especially encouraged.

The total number of registrants for this workshop was 61, and an additional 2 remote, unregistered participants called in for portions of the workshop. Workshop participants were nearly evenly spread across career stages, with 20 Senior Scientists (16+ years experience), 17 Mid-Career Scientists (6-15 years experience), and 23 Early Career Scientists (< 5 years) including 6 Graduate Students & 10 Post-docs.

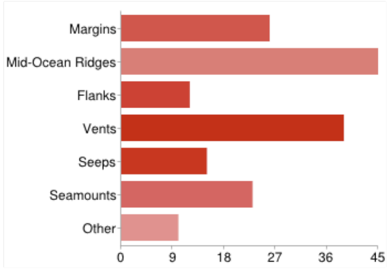
Workshop Participant Demographics


Career Stage

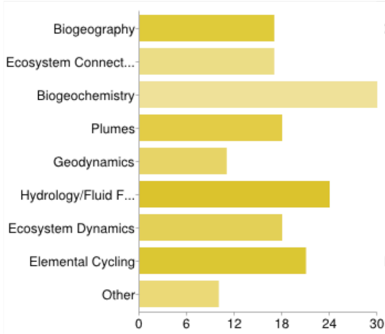
Disciplinary Focus



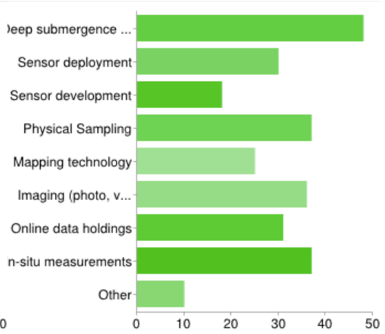
Environmental Focus



Interdisciplinary Focus



Technology Focus



The workshop included several invited speakers including technical and infrastructure perspectives as well as science perspectives:

- Technical and Infrastructure Perspectives
  - Eva Zanzerkia (NSF) -- EarthCube
  - Peter Fox (RPI) -- Geoinformatics and Cyberinfrastructure
  - Vicki Ferrini (LDEO) -- Services provided by the IEDA Data Facility
  - Giora Prioskurowski (UW) -- OOI
  - Dwight Coleman (URI) -- Deep Submergence Telepresence
- Science Perspectives
  - Scott White - (Univ. S. Carolina) -- Geology
  - Breea Govenar (RIC) - Macrobiology
  - Pete Girguis (Harvard) - Microbiology
  - Daniela DiIorio (Univ. Georgia) - Plume Modeling and Fluid Flow

The workshop consisted of several breakout groups and plenary sessions to address both the scientific and technical priorities and challenges within this community. The Science Drivers and Challenges were initially addressed by participants in Discipline-specific breakout groups, and Challenges to Interdisciplinary Science were then addressed by the Interdisciplinary breakout groups. Technical Issues and Challenges were addressed by the Disciplinary groups and Community Next Steps were developed in an open forum plenary session. The 5 disciplinary groups were Geochemistry, Microbiology, Macrobiology, Geology & Geophysics, Physical Oceanography and Crustal Hydrology, and the 5 interdisciplinary groups were Biogeochemistry, Biogeograpy, Geodynamics, Hydrology/Fluid Flow/Plumes, and Ecosystem Dynamics & Connectivity. Each breakout session was followed by a synthesizing open forum plenary session.

In addition to the invited speakers who gave perspectives on both scientific and technical issues, the workshop participants also participated in and heard the results of the Stakeholder alignment survey, given by Dr. Joel Cutcher-Gershenfeld of the University of Illinois. Dr. Cutcher-Gershenfeld's work was complemented by in-person interviews during the workshop. These were conducted by Charlie McElroy, who is a graduate student working with Dr. Cutcher-Gershenfeld. The scientific community's response to both aspects of this work were excellent, with significant participation and interest in the social aspects of our collaborative and interdisciplinary challenges.

SCIENCE ISSUES AND CHALLENGES

1. **IMPORTANT SCIENCE DRIVERS:** Participants identified several high-priority science questions that will be the focus of interdisciplinary efforts during the next 5-15 years.
   - What are the geological/geochemical/physiological/energetic limits of life? What are the boundaries between biological and abiotic control of chemical reactions? How does geochemistry influence microbiology and vice versa? How do we incorporate microbial data into large-scale (global) quantitative geochemical models? How does bioenergetics influence food web dynamics, productivity, energy transfer and nutrient cycles and transform elemental pools between ecosystem compartments? What is the biogeographic, functional and structural distribution of microorganisms and what are the environmental parameters that most influence these distributions? Can these environmental parameters be used as indicators of ecosystem structure and vice versa? How do we define and interpret biomarkers (e.g., paleomicrobiology)? What are the scales of biological responses to disturbance, both natural and anthropogenic and how are these responses reflected in ecosystem connectivity, the relatedness of organisms? Can genetic tools be used to track ecosystem responses to environmental parameters, including adaptation and evolution?
   - What is the architecture of the oceanic lithosphere (including magma processes), and what happens to the plate as it ages from spreading center to subduction zone, as a function of spreading rate, environmental variability, variable crustal architecture? How does plate maturation impact subduction, and what controls the size and cycles of earthquakes in subduction zones? What role does the magma lens play in helping control tectonics/seafloor morphology? Do hot spot/ridge interactions influence the development of oceanic core complexes? What controls the origin, distribution, evolution, and morphology of seafloor features (e.g. seamounts, sulfide mounds), and what is the relationship between these processes/environments on biological communities and mineral resources?
   - What is the role of the deep ocean and subsurface in obtaining a 4D (spatial and temporal) understanding of global chemical and biological reservoirs, fluxes, and energy transfer? Such a perspective would allow us to address such transformative questions as: How does Earth regulate atmospheric CO_2? What are the effects of deep sea biogeochemical processes on modern/ancient global atmospheric chemistry (C,O,S)? What are the relative contributions of biotic and abiotic deep ocean processes to global biogeochemical cycling? How can microbial data be incorporated into large-scale (global) quantitative geochemical models? What are the processes associated with serpentinization, including its diversity, range of environments, and consequences on global elemental cycles? How does the carbonation of peridotites affect global elemental cycles?
   - How do fluids in the subseafloor link thermal, tectonic, seismic, chemical and biological processes in a variety of deep-sea environments? What is the temporal evolution, extent and geometry of fluid flow within oceanic crust? What are the feedbacks between flow and geochemical and geophysical processes? How high within the water column do the fluids go? How does fluid flow effect the transfer of nutrients, energy and heat into habitable zones and what is the role of fluid flow is establishing geochemical gradients and (micro)niches of habitability within the crust?

2. CURRENT CHALLENGES TO HIGH-IMPACT, INTERDISCIPLINARY SCIENCE:

Several themes emerged as consistent challenges faced within/across the involved discipline(s)

- Data integration challenges
  - Communication between more disparate disciplines is lacking in large part because both scientific and funding links are tenuous, but also because there is little history of interaction across these disciplines (e.g., physical oceanographers at biogeography discussions). A key part of future success in interdisciplinary deep ocean studies is encouraging and facilitating communication between disciplines. Progress in this area will subsequently help to overcome the integration of data sets and discipline approaches that are described below.
  - Cross-disciplinary data integration is extremely challenging and true co-registered interdisciplinary data sets are the exception rather than the norm. These challenges stem from issues both at the data collection/management level and with data analysis. For example, data from different relevant disciplines (e.g., biological, geochemical, physical) are not often linked and even the same categories of data are often not comparable in key ways (for example, different molecular samples are processed in different ways and subject to different biases). Few cross-disciplinary data sets exist and deficiencies in acquisition protocols, data quality, and sample metadata make it nearly impossible to link data sets from different disciplines collected on different expeditions. Furthermore, advances in modeling and data analysis techniques are needed to improve cross-disciplinary data integration. For example, merging chemical models with physical or transport models is a science still in its infancy, and new kinds of modeling techniques are needed to integrate heterogeneous data and address interdisciplinary science questions. Within the data management domain, there is both a desperate need for more data in all disciplines and the foreboding challenge of developing tools and platforms (e.g. cloud computing) to handle ever-increasing data volumes and to make data interpolations.
  - To what extent can approaches from one discipline be applied to transform research approaches in other deep sea disciplines? There is significant room for cross pollination of research approaches across disciplines and such activities could be facilitated by categorizing such activities within the scope of broader impacts. In essence, how can you be someone else's broader impact?

- Data acquisition/completeness - particularly with respect to co-registration, and spatial/temporal variability
  - Because deep sea ecosystems are particularly closely linked to geochemical cycles (primary productivity is primarily chemosynthetic and reliant on geochemical fluxes and gradients) there is a need to make spatially and temporally co-registered chemical and biological sampling the norm rather than the exception in deep sea ecosystems. Acquisition of co-registered data is a challenge to current and future deep sea scientists, however access and integration of co-registered data as well as the resolution of legacy disciplinary data into pseudo-co-registered data sets is a challenge that can be addressed by the EarthCurbe Initiative and subsequent data analysis and management tools.
  - The spatial and temporal scales of data collection are very different across disciplines, making interdisciplinary data integration challenging, and many more co-registered, interdisciplinary data sets, collected on comparable scales are needed. For the current data sets, biological occurrence data (in the ocean) are inherently patchy in time and space. Furthermore, many aspects of biogeography are not captured by taxonomic data (such as habitability and energy flows). Additionally, integrating biological data sets

(e.g. spanning the ecosystem from microbe to macrobe) and scaling data sets properly (e.g. measuring specific populations vs. entire communities), could allow us to determine the extent to which specific (keystone) populations drive ecosystem function.

- Understanding ecosystem dynamics requires both the discovery of the required data and synthetic analysis. Therefore the role and challenge of network analysis is to find what you *aren't* looking for that is important – e.g. what data are missing that will make the ecosystem analysis much more robust? Furthermore, adaptive ecosystem behaviors, emergent behaviors, disturbance factors are all challenges to understanding ecosystem function and to developing cyber infrastructure for modeling. Can we develop multi-dimensional datasets to reflect entire ecosystem function?

TECHNICAL INFORMATION/ISSUES/CHALLENGES

1. **Training and Awareness** – Many members of this community recognize that they are supported by existing data management efforts, and clearly stated that they do not want EarthCube to "reinvent the wheel". That said, there is insufficient awareness of and access to existing tools and infrastructure - including data contribution and data discovery tools, open source software, visualization tools, and data analysis systems.
   - New tools need to be developed to improve both data management and data analysis. Particularly, there is a lack of tools that lessen the "burden" of data management and could be embedded in our scientific and daily workflows. New tools that allow for easier and earlier integration of data management activities within the workflow are essential for future data acquisition.
   - There is a large personnel gap in the community between data producers and data managers that could be resolved by facilitating training within the community to lower barriers to available tools and resources.
   - There is significant and well-founded concern that the community lacks sufficient resources for data preparation and that those efforts are not sufficiently recognized and rewarded. While infrastructure for citing data and has been established within several data systems (e.g. Data DOIs), nearly all professional citations continue to be focused exclusively on publications. Much of the hard work of data acquisition, data management, metadata production, data integration is largely unrewarded, lowering the incentive for proper data acquisition and curation and increasing the gap between data scientists and discipline scientists.
2. **Data comparison and integration** -- Datasets are often not fully comparable *because:*
   - Metadata are incomplete and inconsistent;
   - Navigational precision is problematic across deep submergence vehicles. It is essential that exact locations (x, y, z, t) are precisely identified for each sample;
   - Foci differ from project to project. Improving mechanisms for pre-expeditionary communication and developing tools to enhance collaboration (either at particular sites or for particular types of sampling projects) would maximize project utility and drastically increase funding efficiency;
   - Data formats and entry vary from project to project. This can be resolved with either format standardization or, preferably, algorithms that identify and correct for variation in format;
   - There is a lack (or a lack of awareness) of standardized methodologies to document sampling conditions, e.g., consistent definition of time stamps and locations for samples and measurements.
   - Data quality is poorly documented making data use outside the original research group and integration of disparate data sets inconsistent.

3. **Desired Tools**

- Collaborative Tools
    - Tools are needed to facilitate real-time collaboration before, during, and after cruises. These include live ship-to-shore feeds that enhance shore-based participation in sample collection and real-time data analysis. Thus, expedition goals could be dynamic and responsive to real-time data analysis. Furthermore the use of Ancillary Project Letters (APLs) or RAPID-type funding models would allow for interested parties (this could focus on early career scientists) to join expeditions (in person or remotely) to collect co-registered or associated data/samples, thus increasing expedition efficiency. This is important for field-going scientists and modelers alike. This would also lower the barrier for early career scientists to undertake sea-going research by allowing for smaller projects to be funded and completed prior to pursuing larger expedition funding.
    - Mechanisms to better communicate caveats and built in assumptions necessary for interpreting data and models. Models, especially, need to continue to be linked to scientific expertise.
    - "Alert" system that will notify the user of a new data submission of interest. This could be developed to include not only data acquisition updates, but also self-populating personal databases and subsequent data analysis. For example, if one were interested in a particular metabolic functional gene in hydrothermal environments, a search/analysis/model algorithm could regularly self-update and new function gene trees would be the product for the end user. This goes beyond data discovery, but also automates data analysis, allowing scientists to focus on data interpretation.
    - Experimental design, communication/ cooperation with various deep sea and related scientific communities
- Data Documentation Tools
    - There is a lack of tools (desktop, tablets, in the field (ships, ROVs etc)) that facilitate data documentation and capturing metadata that can be used broadly by our community. This is a critical gap that needs to be filled if we are to effectively and efficiently feed content into EarthCube.
    - We also need improved and expanded metadata and standardized metadata templates that easily identify units and commonalities (e.g. when, where (projection, coordinates), how (methods of collection, analysis), experimental design). Furthermore, we need to develop easy tools and simple guidelines for easily capturing metadata contemporaneously at the time of data acquisition.
    - Data quality is inconsistent - EarthCube should include consistent and rigorous mechanisms for objectively documenting and evaluating data quality.
- Visualization and modeling tools
    - Many existing tools require extensive training for effective use or are incomplete. This not only inhibits usage across our community, but also limits our ability to analyze legacy data or integrate and analyze disparate data sets.
    - EarthCube should include a clear and well-organized user interface with a well-documented set of modeling and visualization tools (with training documents) that can be improved or extended in modular form.
    - We need more data integration tools, including tools that easily allow you to merge cross-disciplinary data (different data types) and tools that allow users to look at multiple data sets on a global scale. One idea was: "EarthClip" (*J. Smith*) - Integrated digital (desktop) guidance to help you discover data, contribute data, comment on data quality, etc. (e.g. *"You may also be interested in…"*).

  - Easily accessible interface for using open source tools, without requiring installation on individual computers – cloud based, web page, all encompassing application.
  - Tools needed for interactive figures (3-D) for both processed and raw data.
- Current data sets are enormous and the volume and quantity of data is only going to increase (e.g. HD video is becoming the norm, acoustic datasets, and someday (soon) biologists will be sequencing entire genomes for every organism in a sample). Moving these data sets will be (and is now) an enormous challenge and current solutions are rather antiquated (e.g. we currently ship large hard drives around the globe in order to share data and collaborate on interdisciplinary projects). We need to transition to cloud-based platforms that allow analyses in the cloud with systems that are connected with ultra high bandwidth networks.

4. **Data Curation and Access Issues/Challenges**
   - Relational databases that discern both user interest and intent from search parameters are now common in ecommerce, and could be applied to scientific data searches. For example, when you search for a spatula on Amazon, it shows you a bunch of other spatulas that other users also looked at. Is there a way to have Earthcube know or learn from users about connections between datasets in order to improve data discovery?
   - Access to legacy data is important but is often difficult - EarthCube should include legacy data and/or clear links to legacy data, including ways to objectively evaluate the quality of legacy data. Incorporating legacy data into EarthCube is essential for maximizing its impact in the deep sea science community, however this community will only buy into this platform if there is guaranteed longevity.
   - Lost data sets as well as data sets that don't get pushed into the public domain are not uncommon. We as a community need to continue to be vigilant about data compliance. Can Earthcube make it easier to find and upload data into various databases? Can it be a two way street? Tools that lower the barrier between publications and data upload and curation to data repositories are essential in order to minimize lost data sets and ensure compliance with funding agency requirements for data management.
   - Reducing barriers to access include cross-directorate, cross-agency data linkages ("Data without borders"). This includes NIH-NSF cross communication, potentially combining geological data with 'omics data. Public and private as well as national and international agencies (e.g., ONR, Schmidt, Moore, NOAA, IODP, etc.) support deep sea data acquisition, making data multi-jurisdiction but there is no jurisdiction to the seafloor.
   - Broad-based, interdisciplinary seafloor models and data sets need to be integrated with surface and coastal models, ideally by incorporating all of these in the EarthCube platform.

COMMUNITY NEXT STEPS

1. **List of what your community needs to do next to move forward; how it can use EarthCube to achieve those goals:**

   - We recognize that much of our community is served by existing data management efforts, and recognize that EarthCube should build off these, rather than reinvent them. However, barriers still exist, and we need training to ensure that we can take advantage of existing resources, and to ensure that data are documented and curated accurately and efficiently.
   - As a community, we see very cost-effective rapid solutions to a number of problems that create data management obstacles, but we are unsure what mechanisms might provide funding to address some of the smaller data problems that confront us. While we recognize that there are funding opportunites in EarthCube, it is not clear if any opportunities exist to obtain funding for community-specific projects that could facilitate

inputting data/metadata into the paradigm (e.g. NDSF's Jason Virtual Van upgrades would benefit from cyberinfrastructure input).

- A small subgroup of workshop participants will explore an RCN and/or workshop proposal focused on documenting expedition-based needs. The goal of this effort is to facilitate community consensus to prioritize the needs for improving existing resources for documenting deep submergence field programs - specifically the Jason Virtual Van and Alvin Frame Grabber.
- Tackle Education, Training & Best Practices - PIs, Graduate Students and post-docs need training on how to use available tools for data management, access etc. This can be in coordination with existing data groups that serve our community (e.g. IEDA). NSF should support this training effort. The DEep Submergence Science Committee (DESSC) members who participated in the workshop will pursue this in the context of ongoing early career training efforts.
- Also discussed the concept of a "data wrangler" participating in field programs who is responsible for handling data, and can facilitate contemporaneous data documentation. The role of the data scientist who sits at the intersection of domain science and geoinformatics is rising, but resources are necessary to ensure good data management practices.

EARLY CAREER FEEDBACK - Early career participants conducted a small break-out session of their own to articulate their message to NSF and their perspective on EarthCube:

The dominant concern of our early career participants is related to funding and the bleak job market in academic science research. While they are enthusiastic about their research and the possibilities that EarthCube may enable for them, they are very concerned about career longevity and their ability to pursue cutting edge science in the academic environment.

They also suggested several actionable items we can strive for as a community that would better prepare them for doing better science in a data-enabled world including:

- Small grants for early career scientists to collaborate outside of their institutions.
- Enhanced training opportunities: For example, a data/computer literacy workshop would be broadly useful to early career scientists, and to the deep sea community as a whole.
- Encourage current PIs and mentors to include data/computer literacy into graduate curricula

NSF ACTION ITEMS

The group identified several action items for NSF that would immediately impact this group's ability to not only contribute to, but also be prepared for EarthCube implementation. While we recognize that a majority of the current funding opportunities in EarthCube are focused on developing a cyberinfrastructure to accommodate scientists across the Earth Science disciplines, we believe there are several issues within the deep sea science community that we need to address internally in order for our community to be fully prepared and part of the driving force behind the EarthCube Initiative.

- Develop a RAPID/EAGR-sized funding program (e.g. $50K/award) for discipline/community-specific projects that improve community resources so that they can be incorporated into the EarthCube Initiative. Examples of such community-specific projects include: (i) incorporating legacy data into current data repositories, this could also include rescue of (almost) lost data; (ii) improving data management tools for deep submergence assets (Alvin Frame Grabber, Jason Virtual Van, etc.); (iii) pilot programs for cross-disciplinary scientists to work with data scientists to integrate current data sets and develop small-scale, discipline specific tools that could later be incorporated into and expanded into EarthCube.

- Support data literacy workshops and programs that target both early career and senior scientists. Encouraging young scientists to be not only become data literate, but also to increase their marketability to fill a developing need for discipline trained data scientists within the deep sea community. Furthermore, senior scientists with expedition level responsibilities (e.g. Chief Scientists) need resources and training in data management and curation so that these activities are incorporated into expedition planning and become are integrated early in at-sea work flows.
- Support the incorporation of discipline data scientists into deep sea expeditions. We foresee this would be a multi-tiered program with both support for current data scientists to be integrated into the science expedition team and with support for training of expedition group members in data resource management during pre-expedition planning. We believe it is essential that every science party has a dedicated data scientist to facilitate shipboard data management and enhance data acquisition and documentation, which will serve both the immediate science expedition, and subsequent data users.