From Molecules to Metadata: MIRADA LTERS follow-up working group meeting
Summary Report
March 9th and 10th, 2010
The Ecosystems Center Starr Building, Room 209
The Marine Biological Laboratory in Woods Hole, Massachusetts, USA

Attendees:
Linda Amaral-Zettler, MBL
Melissa Booth, GCE
Hugh Ducklow, PAL
Matt Erickson, PAL
Hap Garritt, PIE
Rafael Guevara, FCE
John Hobbie. ARC
Renzo Kottman, MPI, Bremen
James Laundre, ARC
Liz McCliment, MBL
Phil Neal, MBL
Inigo San Gil, LNO
Tristy Vick, MCM

VTC participants:
John Frisch, CWT
Karen Baker, PAL, CCE
Corinna Gries, NTL
John Porter, VCR
Linda Powell, FCE
Byron Crump, ARC
Wade Sheldon, GCE
Margaret O'Brien, SBC
Elisa Halewood, SBC
Craig Carlson, MCR, SBC
Brian Palenik, CCE
Todd Miller, NTL

The MIRADA-LTERS group held a two-day workshop on the 9th and 10th of March in the Ecosystems Center at the MBL in Woods Hole. We had a total of 25 participants attending with 12 people participating through video teleconferencing. A major goal of the workshop was to bring together data managers and collaborators to validate and finalize sample-associated physical and chemical data for the LTERS contributing to the MIRADA project.

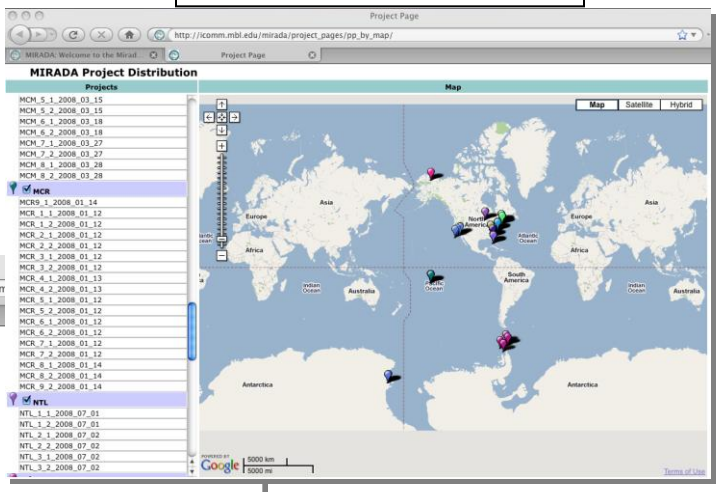
Two weeks prior to the workshop we circulated metadata and contextual data worksheets for each LTER data manager and collaborator to update with new information along with an overview summary of the existing parameters most commonly reported for each group. Each LTER spreadsheet was distributed with questions specific to a given project. Requests included clarification of information for a given parameter or metadata item (i.e. *Is the sampling date correct? What is the latitude and longitude?*), estimates for a parameter given historical data that might be available (i.e. *Does NO2NO3 equal TotN? Can you estimate the pH given past data collections?*) and unit conversion (i.e. *Can you please convert to micromolar?*). Table 1 summarizes the environmental parameters that are available across the MIRADA LTER sites.

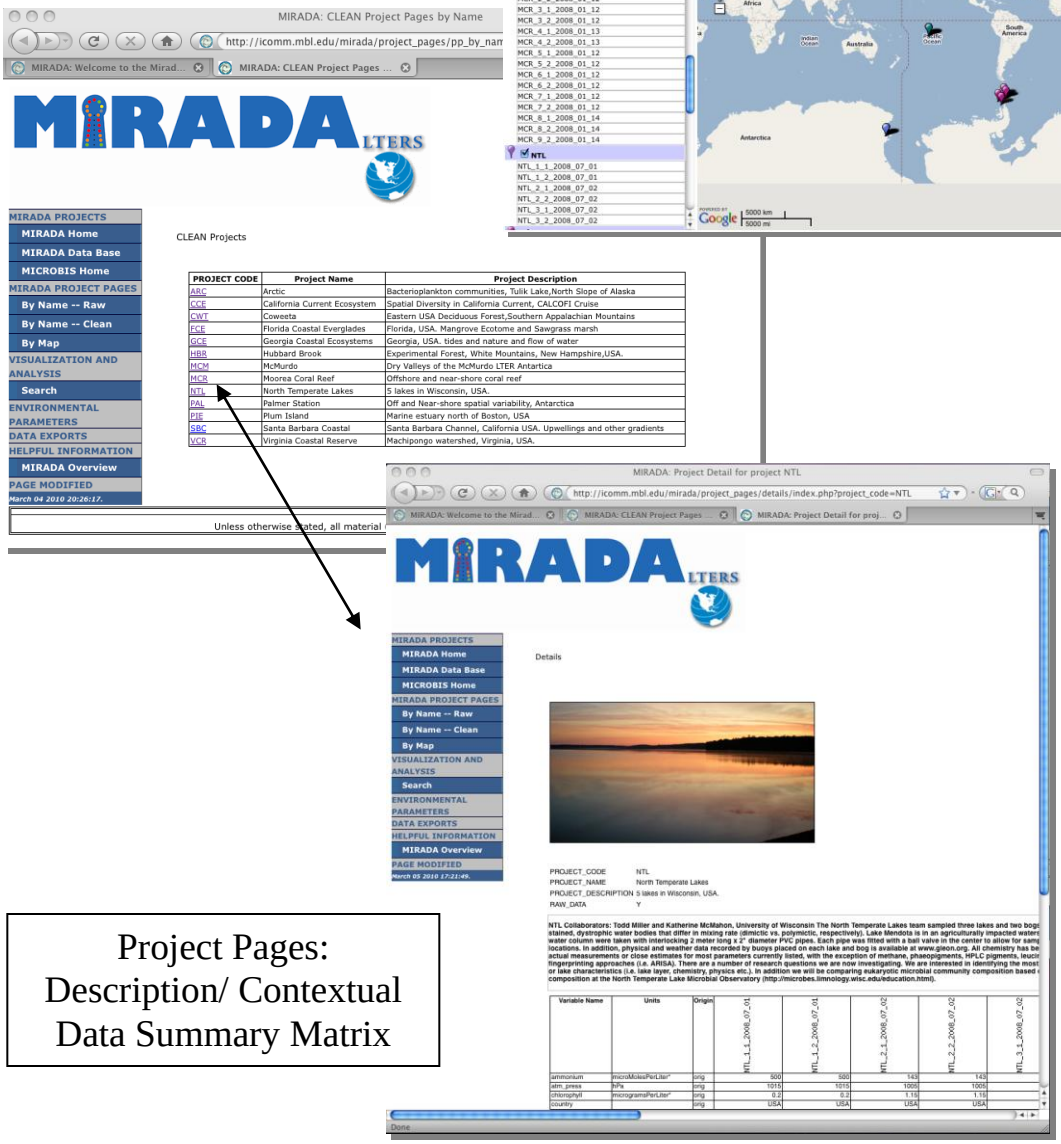
Table 1: A summary of the most commonly reported environmental parameters for the MIRADA participating LTER sites. Highlighted in red are parameters that were recovered either during the two week pre-workshop preparation phase or shortly thereafter. In blue are those parameters that we will strive to obtain from each LTER for the cross-site comparative analyses through estimation via historical data for missing values. The asterisks represent data that are pending. Abbreviations: NH4 – ammonium; CHLA –chlorophyll a; Cond. – conductivity; DOC – dissolved organic carbon; DO – dissolved oxygen; NO2NO3 – nitrite plus nitrate; POC – particulate organic phosphate; PO4 – phosphate; SI – silicate; Temp – temperature; TotN – total nitrogen; TotP – total phosphorus; Lat_Lon – Latitude and Longitude.

Variable	ARC	CCE	CWT	FCE	GCE	HBR	MCM	MCR	NTL	PAL	PIE	SBC	VCR	TOTAL
NH4	X	X	X	X	X	X	X	X	X	*	X	X	X	**12**
CHLA	X	X		X	X		X	X	X	X	X	X	X	**11**
Cond.	X				X	X	X			X	X	X	X	8
DOC	X	*	X	X	X	X	X	X	X	X	X	X		**11**
DO		X		X	X	X	X	X	X	X	X		X	**10**
NO2NO3				X	X		X	X	X	X	X	X	X	**9**
pH	X		X	X	X	X	X		*					**6**
POC		X			X		X	X	X	X	X	X		8
PO4	X	X	X	X	X	X	X	X	*	X	X	X	X	**12**
Salinity	*	X	*	X	X	*	X	X	X	X	X	X	X	**10**
SI	X	X			X	X	X	X	X	X		X		9
Temp	X	X	*	X	X	X	X	X	X	X	X	X	X	**12**
TotN	X		X	X			X		X	X	X	X		8
TotP	X		X	X	X		X		X		X		X	8
Depth	X	X	X	X	X	X	X	X	X	X	X	X	X	13
Date	X	X	X	X	X	X	X	X	X	X	X	X	X	13
Lat_Lon	X	X	X	X	X	X	X	X	X	X	X		X	12

The meeting began with short presentations from each LTER describing the sampling strategy, status of contextual data and questions of interest for the group. We also requested 500 word abstracts and an image that could be posted on the website. Many of these were collected before the meeting and are being used as content for project pages that were generated during the meeting. These pages will be updated periodically with project summary abstracts. These draft MIRADA project pages can be viewed at http://icomm.mbl.edu/microbis/ and by logging into the MIRADA database with the username mirada13 and password vam0s.

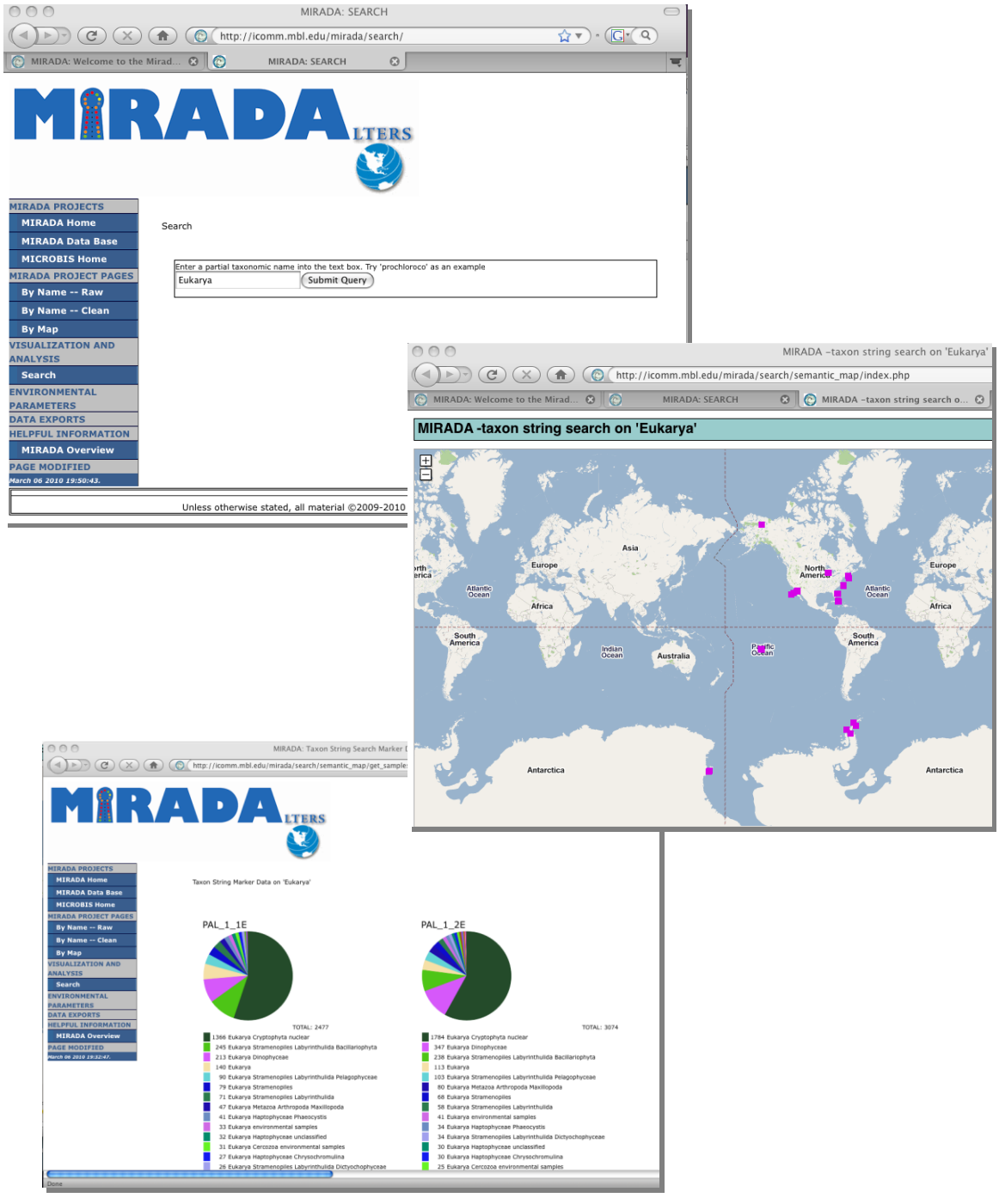
Geographic Location of Datasets





Project Pages: Description/ Contextual Data Summary Matrix

In addition to a sampling rationale and questions of interest, the project pages also provide some simple tools for performing queries across LTER datasets. In the example below, one can perform a semantic search on the term "Eukarya" using the search function available on the MIRADA database front page and recover a map where this taxon occurs and then click on a location to get a taxonomic breakdown of the datasets. These tools have been adapted from the ICoMM MICROBIS website and the efforts of ICoMM IT specialist Phillip Neal.



Another workshop goal was to explore the relationship between MIENS, GCDML and EML. To this end we heard from a colleague from Max Planck Institute, Renzo Kottman who introduced the Minimal Information about an Environmental Sequence (MIENS) concept to the group. The MIENS standard is being adopted by the Genomic Standards Consortium (GSC) (of which Kottman, Amaral-Zettler, and San Gil are members) and is a standard being applied to all of the MIRADA-LTERS datasets. For more details on this please go the GSC website (http://gensc.org/gc_wiki/index.php/MIENS#Latest_News). Renzo also introduced Genomic Contextual Markup Language (GCDML) a type of Extended Markup Language (XML) to implement the MIENS standards that facilitates exchange and integration of genomic data.

Inigo San Gil from the LTER Network Office presented some possible tools based on Drupal that might be adopted by MIRADA LTERS to expand and enrich the existing environmental data associated with the MIRADA datasets. It was agreed that a combination of the existing MICROBIS-style project pages and the implementation of EML to these datasets via a Drupal interface may represent the best path forward. In discussions of contextual data capture associated with the MIRADA molecular data, it was clear that not all the contextual data would be available as EML but that collaborators should strive to convert these data into EML.

The first part of the second day of the meeting was largely dedicated to an introduction of some new tools and manipulation of the datasets developed by members of the MBL Josephine Bay Paul Center informatics team (*Ironing out the wrinkles in the rare biosphere through improved OTU clustering*, Susan M. Huse, David Mark Welch, Hilary G. Morrison, Mitchell L. Sogin Published Online: Mar 11 2010 10:53PM) – namely new clustering and alignment applications to be applied to the datasets. These data analysis improvements will be forthcoming to the MIRADA collaborators. This was followed by an open discussion of possible cross-site comparison target questions. The original plan of separating this session into "fresh water" and "marine" ecosystems was quickly abandoned for a joint session that saw more parallels than divisions. Another realization was that many of the coastal sites actually contain a gradient of fresh to marine stations that were sampled as part of the MIRADA effort. Some suggested interesting target questions that arose during this discussion included the following (in parentheses are the LTER sites that might be considered in addressing these questions):

- Do acidic lake communities show similar patterns in community structure?
- Are there seasonal patterns in microbial communities? (freshwater: MCM, ARC, FCE, PIE; marine; MCR, VCR, PIE, FCE, PAL, SBC)
- Are there community assemblages consistent with freshwater environments?
- Do landscape transfers influence microbial populations in the MIRADA LTERS? (ARC, CWT, PIE, FCE, HBR, GCE,SBC)
- Are landscape-scale patterns visible among MIRADA LTERS?
- What is the extent of integrity in microbial communities in coastal versus open ocean environments? (e.g VCR vs. MCR)

It was also acknowledgement that several of the LTERS sampled along important environmental gradients including:

- Latitude
- Longitude
- Depth
- Salinity
- Oxygen
- Temperature
- Productivity
- Anthropogenic
- Turbidity
- Light

The group also discussed the factors that set the MIRADA cross-site comparative study apart from other efforts including those being lead by the International Census of Marine Microbes. Among the distinguishing factors perhaps the most significant is that we have diversity data from all three domains of life. Other factors acknowledged were the rich contextual data available for all of the datasets, the existing collaborative scientific community, the existence of information infrastructure, the inclusion of freshwater, brackish, coastal and open ocean samples and a replicated sampling design.

The group discussions concluded with recommendations and action items. Several additional recommended parameters were mentioned. For example, the relationship between rain events related to sampling period might be important to consider; % O_2 saturation may be interesting to consider, as well as Photosynthesis: Respiration ratios. Other factors to consider included anthropogenic impacts, land use, permanent cover, landscape modification (e.g dock or bulkhead). Additional recommendations were to consider doing paired testing within sites for temporal comparisons, as well as within-site versus between- site comparisons. The above points were also mentioned as themes for a cross-site publications.

The following additional recommendations and action items were noted:

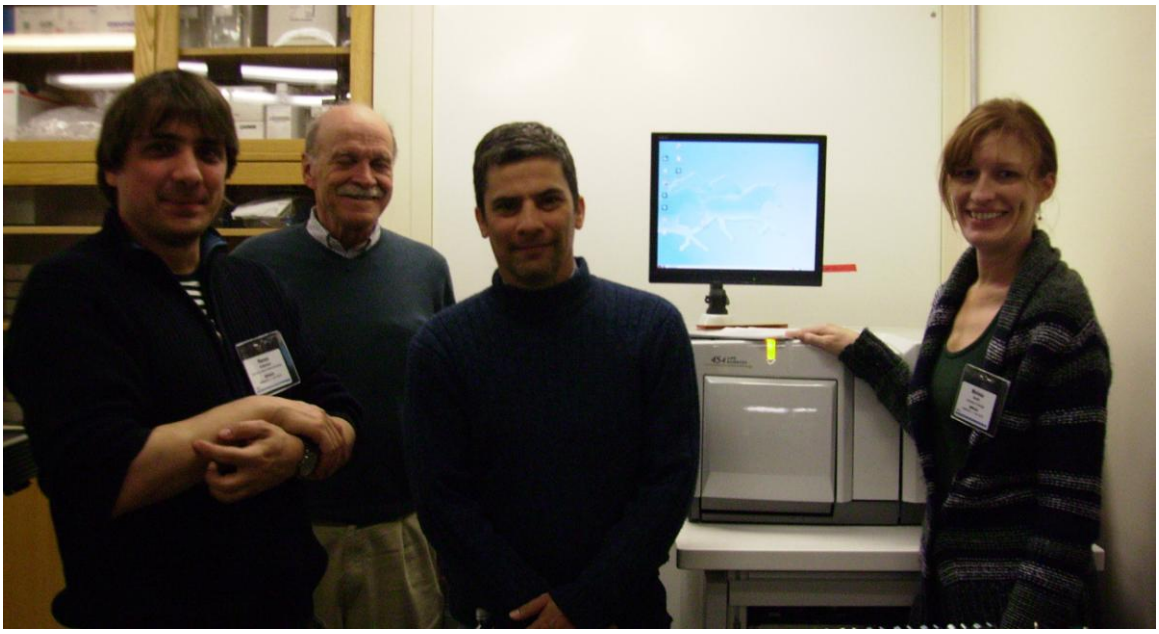
- Determine when contextual data did come from the LTER site and cite the LTER ref id in such cases.
- Provide a link to sources of data.
- Have a metadata field associated with environmental data, at the level of line and individual data point. Point back to original data; attribution and accuracy - create link (DOI-like).
- Provide a location in spreadsheets where one can place an explanation of how the data were processed.
- Provide an EML document for each site that covered post-processing and methods used.
- Find out whether the MIRADA sampling was done in conjunction with LTER sampling.

- Create a second EML document associated with MIRADA dataset.
- Generate a template of EML structure.
- Spell out variable descriptions - look at climedb as an example.
- Consult with the LTER unit working group regarding unit standardization.

The following action items were recommended:

- Inigo to share .csv files of MIRADA data with LTER information managers.
- Confirm nature of when contextual data were collected to ascertain that EML exists.
- Cluster all the MIRADA datasets and provide users with matrices (OTU X dataset).
- Provide CatchAll alpha diversity estimates for all datasets.
- Contact LTER unit working group and decide variable names.
- Renzo to give Hap the controlled vocabulary list.
- Assign an explicit descriptor for each variable name and unit and determine what was measured to convert these units appropriately.
- Go through each spreadsheet and send individual questions.
- Create a roadmap for the rest of project and determine what specific projects will be coming out of this effort.

The workshop concluded with one-on-one discussions with users and meeting organizers and a tour of the MBL Keck sequencing facility.



Appendix 1: Agenda

From Molecules to Metadata: MIRADA LTERS follow-up working group meeting
March 9th and 10th, 2010
The Ecosystems Center Starr Building, Room 209
The Marine Biological Laboratory in Woods Hole, Massachusetts, USA

Monday, March 8, 2010

All day arrival at SWOPE Center, Marine Biological Laboratory

Tuesday, March 9, 2010: Starr Bldg. Room 209

0700: Breakfast
0830: Welcome, logistics, workshop goals **Linda Amaral-Zettler**
0845: **Linda Amaral-Zettler:** The Status of MIRADA-LTERS
0900: **15-minute overviews of Individual MIRADA Sampling Strategies, Status of Metadata and Interesting Questions**
0900: **Liz McCliment** (Moorea Coral Reef, MCR, Virginia Coastal Reserve, VCR & Hubbard Brook, HBR)
0930: **Hugh Ducklow** (Palmer Station, PAL)
0945: **Melissa Booth** (Georgia Coastal Ecosystem, GCE)
1000: **Hap Garritt** (Plum Island, PIE)
1015: Coffee Break
1045: **Rafael Guevara** (Florida Coastal Everglades, FCE)
1100: **Tristy Vick** (McMurdo Dry Valleys, MCM)
1115: **John Hobbie/Byron Crump –** VTC - (Toolik Lake, ARC)
1130: **Craig Carlson –** VTC - (Santa Barbara Coastal, SBC)
1145: **John Frisch** – VTC - (Coweeta, CWT)
1200: **Todd Miller** – VTC - (North Temperate Lakes, NTL)
1215: **Brian Palenik** – VTC - (California Coastal Ecosystem, CCE)
1230: Lunch in SWOPE Private Dining Room
1400: **VTC Session Begins: Linda Amaral-Zettler**: Overview of Metadata/Contextual Data Matrix: Gaps and Challenges
1430: **Inigo San Gil**: MIRADA Project Page Overview, discussion and hands-on demo (Users will create project pages during this segment: They will be asked to bring a site image and sample/project description (500 word max) for the meeting).
1530: **Renzo Kottman**: From Sequence data to Metadata: GCDML, EML and MIENS
1600: Coffee Break
1630 - 1730: "Round Table" Discussion with IMs and users regarding contextual data issues and how to work with historical data to fill in gaps. Assignments for IMs to fill gaps. Outcome should be a matrix of values that can be compared across sites.
VTC Session Ends
1800: Dinner at SWOPE Private Dining Room

Wednesday, March 10, 2010: Starr Bldg. Room 209

0700: Breakfast
0830: **VTC available for remote participation:**
Linda Amaral-Zettler: From Metadata to Molecular Data:
Improvements in Clustering and New Tools for Estimating Alpha Diversity
Visualization and Analysis of Microbial Population Structures - VAMPS
MICROBIS – tools to compare across sites
0930: **Discussion/Question and Answer Session**

1015: Coffee Break
1045: **Break-Out Sessions:** Fresh Water/ Marine groups will separate to come up with a list of target questions and what needs to be done to answer them. What questions can be addressed with the datasets at hand?
1145: **Each group reports back on priority areas of interest for cross-site comparisons: Come up with a task list and what is needed from IMs to facilitate analyses.**
1230: Lunch in SWOPE Private Dining Room
1400: **VTC Session w/ IMs Begins: Review of priority areas for cross-site comparisons between and within fresh water and marine LTERS.**
1530: **VTC Session Ends.**
1530: Coffee Break
1600 - 1800: Individual project question and answer session/ one-on-one assistance as requested.
1600: Optional tour of 454 sequencing facility **(Liz McCliment)**
1800: Dinner at SWOPE Private Dining Room

END OF MEETING

Thursday, March 11, 2010

All participants must check out from SWOPE by 10:00 am. Breakfast provided.