DATA SHARING AND MANAGEMENT PLAN

1. Data description

a. **What data will be collected during this project? How many different data formats are anticipated?**

i. Sequences of DNA and cDNA from microbial and viral communities
   1. Raw image files (e.g. sff files from 454, .fastq files from Illumina).
   2. Fasta format (Raw and quality-trimmed sequence reads, consensus sequence of assembled contigs)
   3. Gene annotation and assemblies

ii. Enumeration data for prokaryotic cells, virus particles, and microbial mortality due to viruses

iii. Geochemical data: nutrient, inorganic, organic, isotope and gas concentrations

iv. Growth and metabolite production rates for isolates and natural assemblages

v. Model input drivers and model output predictions

vi. Engineering drawings for *in situ* incubation instrument and gas-tight samplers.

vii. Seafloor video data from 2012 RSN cruise

b. **When will the data be collected, when will they be entered into electronic databases, and what databases will harbor the data?**

Data collection will occur throughout the project and in several stages. Most data collection will begin in August-September of 2012 with our first field season. Since we are participating in cruises of opportunity funded by NSF/NOAA, we will work with Chief Scientists to deposit relevant data in the Biological and Chemical Oceanography (BCO) Database operated by the NSF BCO-Data Management Office as necessary.

In addition, because GBMF is supporting satellite time for the RSN cruise in 2012, video will be an important data product for our team. The RSN cruise will produce ~10 TB of seafloor ROV video data, ~2 TB of shipboard hand-held video data, and ~1 TB of still images per 3 week cruise. The original footage data will be archived on duplicate high-density tape drives, with one copy stored on campus, and one copy stored in an off-site data repository. Working copies of the raw video and still imagery data will be maintained on multiple redundant arrays of spinning-disk media, from which video pieces are cut and still images edited. Public access to the video and still imagery content is provided through compilations of highlights and select scenes assembled for streaming and download on the OOI website (http://ooi.washington.edu/visions11/video and http://ooi.washington.edu/visions11/gallery) under a Creative Commons license.

In addition to video, data will be collected and entered into databases as follows:

i. Un-annotated sequence data will be released to CAMERA no later than 6 months after their generation.

ii. Annotated sequence data will be generated and made public through the Joint Genome Institute's IMG system, CAMERA, and MBL's VAMPS (vamps.mbl.edu) at the end of years 2 and 3.

iii. Geochemical data will be generated and released to (a) CAMERA, (b) Marine Geoscience Data System: Ridge 2000 data portal (http://www.marinegeo.org/), and (c) VENTDB (www.ventdb.org/) no later than 6 months after its generation and analysis.

iv. Input, output, and computer code for models will be made available through download from laboratory websites once the code is calibrated and stable. Code release will occur throughout the project and all will be released by the end of year 3.

v. If, at the end of this project, it is determined that there is an opportunity for technology transfer of the *in situ* incubation unit to a broader community, then we will consider seeking patents and licensing options to reproduce the technology. Should no patents ultimately be desired or procured, or expected to be procured at the end of the grant term, the engineering drawings will be released via publication in an engineering journal as well as posted on the PMEL websites in (e.g. www.pmel.noaa.gov/edd/hot_fluid_sampler.html). For the gas-tight samplers, we plan to use and/or adapt existing technologies and designs, therefore no new drawings or designs are anticipated.

vi. In addition to the public databases previously mentioned, all final qc'd data products will be entered in an integrated chemical-microbiological database that will be maintained at the MBL or UW. Only project team members will have access to the data through this database during the project term, but this information will be made available through download from laboratory websites by the end of year 3.

c. **Does this project involve organization or analysis of pre-existing data, and what are the data sharing arrangements for these data?**

i. Preexisting 454 16S rDNA sequencing data associated with this project has already been released to the public via VAMPS. Additional functional gene sequence data generated via PCR has been deposited in GenBank.

ii. Preexisting published geochemical data associated with this project has already or will be released to BCO, VENTDB, and the Marine Geoscience Data System.

d. **What are the anticipated data products (e.g. databases, analyses, tools)?**

i. New deep-sea sampling equipment for manipulation, incubation and preservation of diffuse fluids

ii. New integrated chemical-microbiological database of diffuse fluid datasets

iii. New metabolic network model and transport model that incorporates chemical fluxes, bioenergetics, and phylogenetic and functional gene data.

e. **What kinds of metadata will be associated with the data?** See Item 1a for types of data collected, which includes microbial, viral, and geochemical data. Metadata for each sample includes a complete description of the sample location (latitude, longitude, depth, vent name), date/time, temperature, sample number, and sample collection device information.

f. **Who is the owner of the data?** The data will be the property of the principal investigator and co-investigator of the laboratory in which the data was generated.

2. Data management

a. **Where (physically) will the data be stored?** Data will be stored on servers at the Marine Biological Laboratory, University of Washington, PMEL, University of Massachusetts, Amherst and J. Craig Venter Institute. In all cases, appropriate data backup will be practiced. An integrated database for all investigators associated with this project for data sharing and interpretation will be hosted at MBL or UW.

b. **What type of data access or data distribution mechanism and software will be used?** Data distribution will be achieved using common website technologies. For example, for sequencing products produced at the MBL, we have a dedicated sequencing website with an in-house software system that is used to not only distribute software but also perform quality control analysis and statistical tracking of certain types of sequencing data that we commonly produce for internal and external researchers.

c. **Will the location or software for initial data entry differ from the data archive?** We will primarily utilize database servers for data set recording, tracking, and normalization. At the MBL we have 3 high performance database servers that are used for aggregating and working with research data and another 7 databases that are used for dissemination, archiving, and other purposes. In some cases, the data entry interfaces may not be the same as the archiving databases, the distinction of which is generally decided once the project is underway and we are able to more accurately assess the performance, access methods, and other attributes regarding the data sets.

d. **How will metadata be stored, and what provisions will be made to enable metadata searching capability?** Metadata documentation will begin with sample collection. Separate metadata documentation will be created for laboratory experiments. An integrated database for all investigators associated with this project for data sharing, searching, and interpretation will be hosted at the MBL or UW. This database will be made public at the end of year 3. In addition, data will be deposited in public databases as detailed in 1b.

e. **Who will be responsible for entering and maintain data archives, and over what period of time will archives be maintained?** The principal investigator and co-investigators will be responsible for data entry and will ensure archives are maintained through the course of their academic careers.

f. **What data quality controls and assurances will be provided?** The principal investigator and co-investigators will be responsible for data quality control of their individual datasets. Appropriate certified reference materials and/or internal laboratory quality control samples will be analyzed for chemistry.

g. **Who will contribute to the database?** The principal investigator, co-investigators, postdocs, students, technicians, research assistants, and bioinformaticians will contribute to the database.

h. **Will proprietary data be used?** No proprietary data will be used.

3. Data Sharing

a. **Who are the potential data users?** The major users will be investigators from the marine microbial ecology and hydrothermal vent communities. Researchers from other fields, such as terrestrial hot springs, deep biosphere, microbial physiology, astrobiology, ecosystem modeling, and environmental microbiology, are also anticipated.

b. **What is the appropriate timing for release of data to the public or relevant users, and why?** See item 1b. for details on release of data.

c. **When will archived data be openly available to other users?** See item 1b. for details on databases for access to data.

d. **If data from non-GBMF-supported or previous projects are integral to the successful completion of the Grant Purposes, will the non-GBMF-supported and/or pre-existing data also be made freely available?** See item 2c. for handling of pre-existing data. This will be integrated into the final database to be released at the end of Year 3.

e. **How will other users (i.e., beyond the grantee and GBMF) access data and metadata?** See item 1b. for databases for access to all data.

f. **Are the publicly available data in raw form? If not, what treatments have been applied to the data prior to their being released to the public?** See item 1a. for types of sequencing data (raw, processed, and annotated) to be released in databases. For microbial physiology data, raw cell counts and metabolite concentrations will be made available along with the statistical analyses and interpretations of the data following quality control. For geochemical data, final statistically-robust and quality controlled concentration data with associated analytical uncertainties will be released.

g. **How long beyond the grant term will the data be maintained and by whom?** The principal investigator and co-investigators will be responsible for maintaining data archives through the course of their academic careers.

h. **Does the proposed grant include provisions for future hardware upgrades in the event that data is to be stored and maintained well beyond the project period of the Grant?** Not applicable. However, data submitted to national archives will be maintained over long term.

i. **If data analysis tools are to be created as a consequence of the Grant, will a tutorial be available for training of future users of the data, and if so, how can it be accessed?** When we release the final database and models, we will provide tutorials on user access and training via the website.

j. **Will a data sharing agreement be required between outside vendors?** No

k. **Is a Creative Commons type-license appropriate for sharing the data? Why or why not?** By depositing our data in public databases and releasing our own database of integrated data and models at the end of the project, we will be properly sharing our data so I don't think this is necessary. All video and pictures posted on the OOI-RSN website are publicly licensed via Creative-Commons.

l. **How will appropriate attribution to the data provider by provided?** Data will be appropriately attributed through authorship of publications and public database entries.

m. **Do you anticipate publishing a "Data Release Paper" for referencing and sharing the data?** This is yet to be decided. In the event that we release a large geochemical dataset to VENTDB, for example, we would consider submitting a data report type publication to inform the community and describe the database and its utility for modeling.

n. **Other items.** See section 1(b)(v) for data release plans related to design of the in situ incubation units.