16 April 2013

KAY DANIEL BIDLE

MMI Investigator Award **Data Sharing and Management Plan***

Data Description

What data will be collected during this project?

- Biological data
  - Physiology (cell abundance, growth dynamics, growth rates, chlorophyll, etc.)
  - Photophysiology and biophysical (e.g. F_v/F_m, σ, linear & cyclic electron flow)
  - Bioptical (scattering, absorption spectra, etc.)
  - Virus abundances and production rates
  - Enzyme hydrolysis rates (e.g. caspase activities, etc.)
  - Transparent exopolymer particles, aggregation
  - Cellular nitric oxide production
  - Cellular reactive oxygen species (ROS)
  - Membrane integrity staining
  - Microscopy images
- Biochemical Data
  - Western blot analyses
  - Enzyme activities
  - Structural protein analysis
  - GFP expression/reporter analysis
- Molecular Data
  - DNA gene sequences
  - RNA transcript identities and abundances
  - Bioinformatic analyses
  - Molecular biology constructs
- Flow Cytometry Data
  - Scatter plots of chlorophyll, side-scatter, forward scatter, fluorescence
  - Cell sorting data
- Chemical analysis of putative infochemical/biomarker molecules
  - Distributions and concentrations
  - Chromatograms
  - Structural analysis
- Field data
  - Biological, biochemical, molecular and flow cytometry data from the field
  - Nutrient concentrations
  - Standard physical and chemical oceanographic data

How many different data formats are anticipated? Please list formats.

- *Biological, biochemical, molecular, flow cytometry, and raw field data* will be collected in near-real time, and then manually entered into a common excel spreadsheet for each culture, experiment, and/or cruise.
- *Chemical data of putative infochemicals* obtained from instruments (i.e. mass spectra, chromatograms, subtractive mass spectra/chromatograms) will be initially generated in proprietary formats via the software that runs each of instruments. These data will be moved to the common excel spreadsheet for each culture, experiment, or cruise.
- *Finalized datasets* will be obtained through processing, manipulations, and quality control of raw data in excel spreadsheets. *Finalized datasets* will be of fully publishable quality and are meant to be used by the PI and the public without any further necessary data manipulations.

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

- *Biological, flow cytometry and some biochemical data* will be collected in near-real time, with *biochemical, molecular, and chemical data* lagging a few weeks/months behind
- Experimental *raw data* will be of little use to anyone aside from the investigator that collected it, and thus we will not establish an open database for this information.
- Field data will have different frequency of collection depending on its nature. Standard physical and chemical oceanography metadata will be collected in near-real time and will be deposited into the 'Rolling Deck to Repository' (R2R; http://www.rvdata.us). Collection/generation of nutrient, biological, molecular, and flow cytometry data from the field will lag behind by several months. Ultimately, all data related to fieldwork will be posted on the publicly-accessible BCO-DMO website (http://bcodmo.org)
- *Finalized datasets* will be posted to an open database on the project website, along with a list of high-level category descriptions of ongoing and completed experiments. This should enhance opportunities for collaboration.
- *Finalized molecular datasets* will be deposited into the NCBI Genbank (http://www.ncbi.nlm.nih.gov) and CAMERA (http://camera.calit2.net) databases.

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

- N/A

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

- The database of *finalized datasets* will be the primary immediate data products. However, we expect that even our *finalized datasets* will be of little use to the public, because it is the experimental and environmental context that gives the data meaning.
- As far as the broader scientific community is concerned, the primary data products will be peer-reviewed publications, with open-access.

What kinds of metadata will be associated with the data?

- Molecular metadata (experimental conditions, etc.) will be collected and included in the NCBI and CAMERA databases. Where appropriate, molecular metadata will interface and conform with community-driven standards/methods for capturing and exchanging genomic metadata as developed by The Genomic Standards Consortium (GSC; www.gensc.org)
- Field metadata from our research cruises, such as weather parameters, depth, position, and standard physical oceanographic data will be collected and archived in the R2R and BCO-DMO databases.

Who is the owner of the data?

- The quality of the *finalized datasets* generated in this project will be the responsibility of PI (and/or PIs in charge of lab group(s) of related MMI projects; see note below).
- Once posted, *finalized datasets* will be public and will not be "owned" by any PI.
- PIs will retain ownership of *raw data* generated by their groups.

Data Management

Where (physically) will the data be stored?

- *Finalized datasets* will be stored on Rutgers' IMCS servers and posted to the public on the PIs project website.

What type of data access or data distribution mechanism and software will be used?

- *Finalized datasets* will be posted as excel spreadsheets.

Will the location or software for initial data entry differ from the data archive?

- N/A

How will metadata be stored, and what provisions will be made to enable metadata-searching capability?

- Field metadata will posted on the fully searchable BCO-DMO website.
- Molecular metadata will be stored in the NCBI and CAMERA databases

Who will be responsible for entering and maintaining data archives, and over what period of time will archives be maintained?

- PI Bidle (with the help of senior lab personnel) will responsible for seeing that *finalized datasets* are archived at Rutgers, and maintained for at least three years after the grant term has ended.

What data quality controls and assurances will be provided?

- Our project will yield myriad types of data, each with their own suite of uncertainties. The only QC/QA check will be at the PI (or senior personnel) level, when *finalized datasets* are submitted for posting.

Who will contribute to the database?

- The PI (and associated project personnel) will be the only contributors of *finalized datasets* to the database. They will collect, assemble, and QC/QA *finalized datasets* from the members of their respective lab groups.

Will proprietary data be used? If so, describe the permissions obtained to use the data.

- N/A

Data sharing

Who are the potential data users?

- The *finalized datasets* will be open to the public. As stated above, it is expected that only the PI and associated personnel will be the primary users these data.
- It is also anticipated that both *raw* and *finalized* datasets will be of interest and useful to lab group(s) (PIs and associated personnel) of related MMI projects; see note below).
- Peer-reviewed publications will be used by the broader scientific community.

What is the appropriate timing for release of data to the public or relevant users, and why?

- The *finalized datasets* will be posted as soon as they are complete. It is doubtful that anyone outside of our project will be interested in these, so there is likely no need for a time-consuming embargo for scientific purposes, unless they are part of a submitted/published graduate thesis and require sufficient time to work up for publication in the primary literature.
- Policies of the PI's institution may require that the commercial value of *finalized datasets* be secured prior to posting (e.g. preliminary patent application) prior to posting. At no point should this process lead to lengthy delays in posting (Bidle has already worked together with collaborating PIs, including those on a related MMI project, on joint patent applications).
- Where appropriate, *raw* molecular 'omic' data will be posted on the aforementioned databases.
- Release/posting of *raw* and *finalized* datasets will follow by the MMI's annual data release checkpoints. The PI will disclose (1) what has been published / patented; (2) what has not yet been published / patented but that is expected soon; (3) what will soon be published but can be made public in advance anyways; and (4) what will not likely be published/patented and is being made public. Based on these checkpoints, the PI (along with lab personnel and relevant collaborators) will have 6-18 months to roost on data before it goes public. These checkpoints will provide timeframes for (2) with the goal of getting data/knowledge into categories (1), (3), and (4) as soon as possible.

When will archived data be openly available to other users?

- *Finalized datasets* will be available immediately upon posting

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?

- Yes, but only in the form of *finalized datasets*.

How will other users (i.e., beyond the grantee and GBMF) access data and metadata?

- They will be publically available.

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?

- All publically available data will be in the form of *finalized datasets*.

How long beyond the grant term will the data be maintained and by whom?

- The *finalized datasets* will be maintained on the project website for three years after the grant term has ended.
- *Metadata (both field and molecular)* will be maintained indefinitely by the aforementioned databases

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?

- No.

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?

- N/A.

Will a data sharing agreement be required between outside vendors? If so, a brief description of the agreement needs to be provided in the grant proposal.

- N/A.

Is a Creative Commons type-license appropriate for sharing the data? Why or why not?

- No. *Finalized datasets* should be freely available to the public without any need for licensing of any type. The complexity of the scientific tasks and resultant *finalized datasets* are sufficient protection against anyone using our data in a scientifically meaningful manner.

How will appropriate attribution to the data provider be provided?

- Where relevant datasets are produced as part of collaborations (as with related MMI projects), the affiliated PIs have worked together extensively and share authorship on a number of occasions; consequently, a formal protocol is unnecessary.

Do you anticipate publishing a "Data Release Paper" for referencing and sharing the data?

- No

---

* Note: Research activities and components of this Data Sharing and Management Plan will interface closely with activities outlined in MMI Grant Project #3301 ("*Infochemical control of microbial interactions and nutrient cycling in blooms of diatoms and coccolithophores*) for which Bidle is a co-PI.