Deepwater Horizon Oil Spill (DWHOS)
Water Column Technical Working Group

Image Data Processing Plan: *IS*IIS

Principal Investigator: Dr. Robert K. Cowen, UM/RSMAS

August 9, 2012

Prepared by:
Robert K. Cowen (UM/RSMAS), Eileen Graham (ASA), Deborah French McCay (ASA) on behalf of the Trustees
Reviewed by:
NOAA: Dan Hahn (NOAA), John Quinlan (NOAA)
Louisiana: Amanda Vincent (LDEQ) and LDWF Office of Fisheries
BP/Cardno ENTRIX: William Graeber and Jeffery Simms

Period of Performance
One year from start date; estimated start date: July 1, 2012.

1.0 Objectives and Approach
This plan describes the image analysis (IA) procedures specific to imagery data taken as part of the NRDA Water Column Technical Working Group's sampling effort using the *In Situ* Ichthyoplankton Imaging System (*IS*IIS) during the summer of 2011 (Principal Investigator (PI): Dr. Robert K. Cowen, University of Miami, Rosenstiel School of Marine and Atmospheric Science, UM/RSMAS). Other work plans describing analysis of data from Shadowed Image Particle Profiling and Evaluation Recorder (SIPPER, PI: Andrew Remsen); Holocam, Digital Autonomous Video Plankton Recorder (DAVPR), and Video Plankton Recorder (VPRII, PI: Cabell Davis); and FlowCAM and ZOOSCAN (PI: Malinda Sutor) involve similar analytical procedures and the processing for these data will be included in separate processing plans.

As part of the NOAA/NRDA Water Column TWG sampling effort, digital imagery was collected during three separate cruises on board the NOAA ship *McArthur II*. During these cruises, the field work involved collection of raw digital imagery data onto archival disk drives for later, shore-based analysis. For *IS*IIS, the imagery data uploads as a continual image which is then parsed into consecutive images that are individually saved with time stamps into stacked files (300 images per file). These archived images require post-processing to: i) extract regions (segmentation) of each image that correspond to possible organisms (Regions of Interest – ROIs); ii) classify each ROI according to a training library of known image classes (e.g., larval fish, chaetognaths, etc.); iii) measure and count each class; and iv) input and store images and all environmental and imagery data into object-relational database structure.

This plan specifically covers these processing steps:

1. Develop Object-Relational database and appropriate user-interfaces (UI).

2. Optimize (recompile) segmentation and classification codes for High Performance Computing (HPC) and General Purpose computing on Graphic Processing Units (GPGPU) environments.
3. Run segmentation and classification of all imagery on HPC/GPGPU machines.
    a. Extract ROIs of each image that correspond to possible organisms.
    b. Classify each ROI according to a training library of known image classes (e.g., larval fish, chaetognaths, etc.).
    c. Measure and count each class.
4. Upload all resulting IA data, including ROI images, counts and sizes of each classified ROI, and time stamped environmental data into object-relational database.

The end data products will consist of: the extracted sub-images (in a common digital photo format), a database of image information (location, classification, measurements, etc.), summary tables and figures, and a narrative description of the data. Determination of the geographic and depth distribution of these data will also be carried out.

2.0 Methodology

2.1 Background

The *In Situ* Ichthyoplankton Imaging System (*IS*IIS) is a high resolution imaging system designed to image sufficient volumes of water to quantify rare mesoplankton such as larval fish and associated plankters (e.g., ctenophores, larvaceans, chaetognaths, hydrozoans, scyphozoans, pteropods, decapod larvae) *in situ* (Cowen and Guigand, 2008). The system is capable of imaging up to 140 liters per second, yet with pixel resolution of 68 μm, yielding particle resolution down to ~500 μm. The *IS*IIS is also instrumented with a set of environmental sensors including: conductivity, temperature, depth (CTD); photosynthetically active radiation (PAR); dissolved oxygen (O_2); and fluorometry. All data is passed up via fiber optic cable to an onboard computer where it is time-stamped for cross-referencing and saved on a high speed disk array (see Attachment 1 for more details). Though a relatively new instrument, *IS*IIS has been field tested in a variety of environments including the Florida Straits, NE Middle Atlantic Bight, Gulf of Maine, Monterey Bay (CA), and the Southern California Bight. Several papers are in review/press following from these various deployments (e.g. Cowen et al., in review; McClatchie et al., in press).

During the summer of 2011, three cruises were conducted onboard the *McArthur II* using *IS*IIS as the primary sampling platform (Table 1). The primary activity of these cruises (3 excursions under one workplan) was to collect plankton image data using *IS*IIS in both open ocean (deepwater and slope, > 200 m) and shallow water (shelf, < 200m) environments within ~ 100 km of the wellhead (see Figure 1) to define the spatial distribution and abundance of mesoplanktonic organisms and assist in understanding the potential exposure of mesoplanktonic organisms to oil. In particular, image data was collected in order to:

1. Identify the broad to finescale vertical and spatial patchiness of mesoplankton from open ocean onto and across shelf environments,
2. Identify the finescale horizontal patchiness of shallow mesoplankton within both open ocean and shelf environments, and
3. Identify the finescale diel vertical distribution of mesoplankton in both open ocean and shelf environments.

Table 1. Cruises where *IS*IIS data were collected.

Cruise	Date	Approximate Location
McArthur II 2	June 12 – June 30, 2011	Cross-shelf, within ~100 km of wellhead
McArthur II 4	July 18 – July 31, 2011	Cross-shelf, within ~100 km of wellhead
McArthur II 5	August 21 – September 2, 2011	Cross-shelf, within ~100 km of wellhead

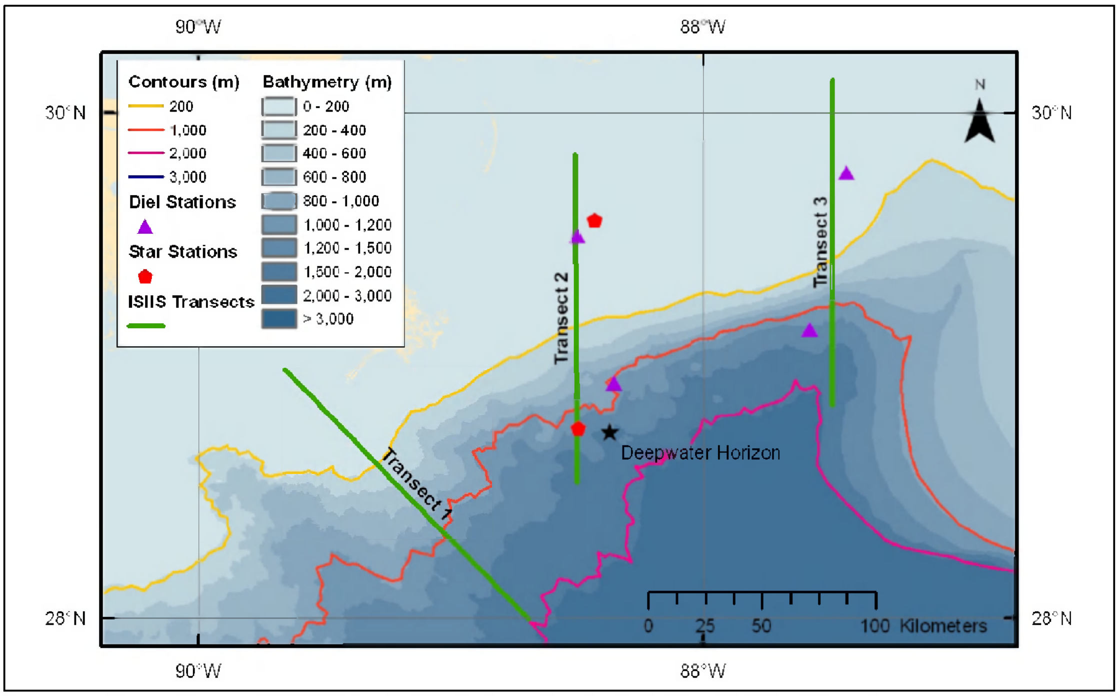

Figure 1. Locations of NRDA *IS*IIS data collected in the Gulf of Mexico summer of 2011.

For the duration of sampling over the three cruises, we collected a total of nearly 49 terabytes (TB) of image data, which comprised a total of ~10,700,500 individual (though consecutive) images requiring image analysis (see Table 1 in Attachment 2).

2.2 Specific work tasks:

There are several steps required to complete the image analysis of the *IS*IIS imagery, and save the resultant data into a user-friendly and accessible database for further evaluation (for more detail on these procedures, see Attachment 2). Moreover, specific steps must also be taken to ensure proper QA/QC and for inter-comparison with digital image data of Cabell Davis (WHOI), Andrew Remsen (USF) and Malinda Sutor (LSU).

1. *Develop Object-Relational database and appropriate user-interfaces (UI).* Given the very large data load (49 TB of raw imagery) to be analyzed (for ROIs and classification), we will develop an object-relational database engine (e.g., *PostgreSQL*) for all data storage capable of receiving direct download from analysis programs. This database will be developed on servers at the UM Center for Computational Sciences (CCS) with password protected access.

We will also develop, install and test the supporting utilities needed for server-support of the database (e.g., Apache v2, PHP, Drupal).

2. *Optimize (recompile) segmentation and classification codes for HPC and GPGPU environments.* Again, due to the large data load, all steps of the image analysis will require access to high speed computing platforms (whether via CPU or GPU multiple core computers). We will optimize our segmentation and classification algorithms for parallel HPC and GPU computing environments. The software will control all aspects of data reading, analysis, and recording (e.g., unstacking image files, maintaining time stamp in all filenames of imagery and supporting environmental data, segmentation and classification).

3. *Run segmentation and classification of all imagery on HPC/GPGPU machines.* To maximize computing power directed to the actual image analysis (once all programming is optimized for relevant computing platform), we will process on both (HPC and GPGPU) systems. Given the very large number of image files, we will have access to hundreds of cores on the HPC machines within CCS, as well as a state-of-the-art GPGPU system.

a) Extract ROIs of each image that correspond to possible organisms. ROIs typically contain only one object of interest but multiple objects can sometimes be found in one ROI. ROIs with more than one object, or overlapping objects, are classified in a "multiple objects" category or as "unknown". Both of these categories require manual review to properly classify all objects in the ROI.
b) Classify each ROI according to a training library of known image classes (e.g., larval fish, chaetognaths, etc.). Prior to classification, classification libraries (or training image sets) will be developed by the Cowen lab. These libraries will also be made available to the other digital imaging labs (listed above) for both their own potential use, and QA/QC activities discussed below.
c) Measure and count each class. The size of objects will be measured digitally through a conversion of pixel to distance calibration. With these calibrations, any particle/organism can be measured accurately by direct counts of pixels along the longest axis. As these values are automatically taken by the image analysis software, particle sizing can be completed on any selected class. Manual evaluation is required to assess precision issues related to object orientation (i.e., extent to which object is orthogonal to field of view). Through subsampling classified images, orientation can be assessed and correction factors derived/applied, if necessary. In addition, the total number of images captured for a given category (e.g., suspended particulate matter (SPM)) per unit time is converted to number of objects per unit volume using the calibrated image volume and sampling rate for each instrument (volume per unit time).

4. *Upload all resulting IA data, including ROI images, counts and sizes of each classified ROI, and time stamped environmental data, into object-relational database.* Per item 2 above, all data will be uploaded into the object-oriented database developed for this project. This database will include not only the image data (raw and segmented images), but also all of the time-stamped linked environmental data, and the classification data (including count and size data if taken).

a) Time-match each ROI with concurrently collected and deployment data. Once images have been classified, they can be time-matched with CTD data to obtain location and depth information for each image.

3.0 Data Processing Prioritization and Schedule

The *IS*IIS data sets will be processed sequentially. The June cruise will be processed first (i.e., priority level 1), followed by the July cruise (priority level 2), and finally the August cruise (priority level 3). The schedule for completion of work, estimated based on the overall schedule, is for the June cruise data to be delivered in month 7, the July cruise in month 9, and the August cruise in month 10. Data will be distributed after it has gone through the QA/QC process described below and in accordance with the distribution procedures, also described below.

4.0 Quality Assurance and Control

Quality assurance of data products will be achieved via a phased QA/QC process to be implemented across the NRDA plankton/particle imaging PIs (Cabell Davis (Holocam, DAVPR, VPRII), Andrew Remsen (SIPPER), Robert Cowen (ISIIS), and Malinda Sutor (FlowCAM, ZOOSCAN)). The first phase of QA/QC will involve the sharing of computerized classification methods among the PIs working with digital image analyses of plankton and other classification categories. This activity will assess the degree to which the PIs are classifying imaged objects to a similar level (e.g., SPM and major taxa of plankton such as copepods, etc.). If needed, taxonomic resolution will be adjusted to achieve the best possible agreement between image classification systems, while recognizing the differences in size classes evaluated and discrimination capabilities such that identification is to the lowest possible taxon for each instrument. It is envisioned, but not required, that the initial discussion between the PIs will occur at an in-person meeting. Following this initial cross-coordination, communication is expected to occur on an as-needed basis, as decided by the PIs and the Trustees. A report will be compiled and provided to all parties to this agreement which will document the common classification schemes that provide the basis for integration across the listed imaging systems. The report will also provide a comparison of the level of taxonomic resolution between the instruments/systems used. If any major updates or changes are made to the classification systems as the data processing progresses, an amended report will be developed and distributed.

A second phase of QA/QC will be performed at the software level. This includes identification of training images, the creation of training image sets, and conducting routine manual checks of image processing software. These three tasks will be carried out by Robert Cowen and technicians during the processing of each data set. Trained technicians will perform the manual identifications, and Dr. Cowen will check their work by identifying a randomly selected subset of at least 5% of the images. Additionally, detection and false positive rates will be determined via confusion matrices generated when building the automatic classifiers from sets of training images (Hu and Davis, 2006). A confusion matrix compares the classifications made by a human expert (columns) to those made by the computer (rows). The accuracy, precision and comparability of the identifications varies with taxonomic group. By using equal numbers of images from the training set for each category classified, classification accuracy is quantified by the confusion matrices (i.e., the confusion matrices show the proportion of false positives, true positives, false negatives, and true negatives). Abundances (e.g., number of organisms per liter) are corrected using the information from the confusion matrix. Abundances generated by this automated method are similar to those based on human sorted images (Hu and Davis, 2006). Confusion matrices and associated abundance corrections (Hu and Davis, 2006) will be shared among plankton/particle imaging PIs to ensure the different software systems are performing similarly, and will be included in the official data record released to the Trustees and BP. Any identifications completed for which there is no automated process (e.g. ROIs with multiple organisms) will also be subject to a 5% check of the images by Dr. Cowen.

A third phase of QA/QC will involve an independent cross-check between the principal investigators working with imaging systems (Davis, Remsen, Cowen, and Sutor). At least 50 images will be randomly selected from each taxonomic group and sent to at least one other PI for their independent identification by manual inspection. This will serve as a quality control check for Dr. Cowen's system and will be part of an integrated system of QA/QC checks across the plankton program. The PIs will independently identify images but will also confer with each other about classifications. As necessary, Cowen will also make adjustments to software, as possible, to ensure maximum agreement among investigators.

The final phase of QA/QC will be to determine that the data are reported in geographically correct ranges, as compared to field-recorded vessel and station locations for the imagery collections. This task will be performed by the NOAA NRDA Data Management Team after the lab data for a priority level is received and before the data are released to the Trustees and BP. Since the whole process is based on electronic data, transcription from paper data sheets to electronic media is not anticipated. However, if transcription does occur, a cross-check of 100% of all transcriptions will be conducted by someone other than the person who completed the original transcription.

5.0 Distribution of Results

Upon completion of the data processing and inter-PI QA/QC procedures (QA/QC phases (1) – (3) as identified above) at each priority level, the principal investigator (Dr. Robert Cowen) will deliver all products generated as part of this work plan (e.g., processed data sets) to the NOAA NRDA Data Management Team. Once the Data Management Team has completed their QA/QC review of the station metadata for that priority level (the final phase above), the data and metadata for that priority level will be made available to the parties of this agreement by means appropriate to the data type as determined by the NOAA NRDA Data Management Team. NOAA and the Louisiana Oil Spill Coordinator's Office (LOSCO) on behalf of the State of Louisiana and BP (or Cardno ENTRIX on behalf of BP) will be alerted when these data become available.

In the interest of maintaining one consistent data set for use by all parties, only the verified and validated data set made available to the parties by the NOAA NRDA Data Management Team shall be considered the consensus data set. In order to ensure reliability of the consensus data and full review by the parties, no party shall publish consensus data until 14 days after such data have been made available to the parties. Any questions raised about the consensus data set, as it was made available to the parties by the NOAA NRDA Data Management Team, shall be handled consistent with the procedures in Section 7.2 of the Deepwater Horizon NRDA Analytical Quality Assurance Plan.

- The trustees and BP shall each designate an individual responsible for raising questions, if any, on the consensus data set.
- If questions are raised, the two designated individuals will meet to determine the source of the difference and resolve.
- The questions raised and their resolution shall be distributed to all parties.
- No changes to the consensus data set will be made if the differences are considered immaterial by both designated individuals, acting on behalf of the parties.
- If the parties agree that changes to the dataset should be made, the dataset will be updated in accordance with the resolution and reposted with a notation that the dataset has been revised.
- If the designated individuals do not agree on how to resolve the difference concerning the consensus data set, the designated individuals shall request assistance from the Assessment Managers for the trustees and BP.

6.0 Retention of Materials

All imagery, data and other products developed pursuant to this plan will be retained, along with any changes in processing software or results. All of this information will be maintained during all review steps in the process and stored in secure locations under Trustee control and will be provided to all parties as part of the data release process.

All materials associated with the collection or analysis of samples under these protocols or pursuant to any approved work plan, including any remains of samples and, including remains of extracts created during or remaining after analytical testing, must be preserved and disposed of in accordance with the preservation and disposal requirements set forth in Pretrial Orders ("PTOs") # 1, # 30, #35, # 37, #39 and #43 and any other applicable Court Orders governing tangible items that are or may be issued in MDL No. 2179 IN RE: Oil Spill by the Oil Rig "DEEPWATER HORIZON" (E.D. LA 2010). Destructive analytical testing of oil, dispersant or sediment samples may only be conducted in accordance with PTO # 37, paragraph 11, and PTO # 39, paragraph 11. Circumstances and procedures governing preservation and disposal of sample materials by the trustees must be set forth in a written protocol that is approved by the state or federal agency whose employees or contractors are in possession or control of such materials and must comply with the provisions of PTOs # 1, # 30, # 35, 37, #39 and #43.

7.0 Budgeting

The Parties acknowledge that this budget is an estimate, and that actual costs may prove to be higher due to a number of potential factors. The costs indicated in Budget Chart # 1 below and any additional reasonable costs within the scope of this workplan that may arise shall be reimbursed by BP upon receipt of written invoices submitted by the Trustees. The Trustees will make a good faith effort to notify BP in advance of any such increased costs.

Budget Chart #1.

Task	Base	IDC	Total	Estimated Time Frame
Project oversight – coordination among CCS and UM teams; communication with NRDA trustees; project reporting	[redacted]	[redacted]	110K	Equal portion per month

Task 1: Develop Repositories (plus raw image data storage data storage)	[redacted]	[redacted]	126.6K	Completion by end of month 3
Task 2: Optimization of segmentation/classification codes	[redacted]	[redacted]	239.4K	Completion by end of month 5
Task 3: Run Segmentation/classification operations	[redacted]	[redacted]	208.8K	Completion by end of month 10
Task 4: Populate DB – all raw data; metadata; environmental data; processed image data; develop full documentation	[redacted]	[redacted]	104.3K	Completion by end of period of performance
Task 5: QA/QC	[redacted]	[redacted]	111.7K	Completion by end of period of performance
Estimated Total	[redacted]	[redacted]	**$900,900**	

8.0 References

Cowen R.K., and C.M. Guigand. 2008. Ichthyoplankton Imaging System (*IS*IIS): system design and preliminary results. *Limnol. and Oceanogr. Methods* 6: 126-132.

Cowen, R.K., A. Greer, C.M. Guigand, J.A. Hare, D.E. Richardson, H. Walsh. In review. Evaluation of the *In situ* Ichthyoplankton Imaging System (*IS*IIS): comparison to traditional (Bongo) net sampling. *Fishery Bull.*

Hu, Q. and C.S. Davis. 2006. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. *Mar. Ecol. Prog. Ser.*, 306, 51-61.

McClatchie, S., R.K. Cowen, K. Nieto, A. Greer, J. Luo, C.M. Guigand, D. Demer, D. Griffith, D. Rudnick. In press. Resolution of fine biological structure including small narcomedusae across a front in the Southern California Bight. *J. Geophysical Research.*

9.0 Attachments

Attachment 1. *IS*IIS Specifications
Attachment 2. *IS*IIS Processing Protocol

Deepwater Horizon Oil Spill (DWHOS)
Water Column Technical Working Group

Image Data Processing Plan: *IS*IIS

Principal Investigator: Dr. Robert Cowen, UM/RSMAS

August 9, 2012

Approvals

Approval of this work plan is for the purposes of processing imaging data for the Natural Resource Damage Assessment. Parties each reserve its right to produce its own independent interpretation and analysis of any data processed pursuant to this work plan.

BP Approval Joyce Miley Printed Name [signature] Signature 11/5/12 Date

Federal Trustee Approval ________ Printed Name ________ Signature ________ Date

Louisiana Approval KAROLIEN DEBUSSCHERE Printed Name [signature] Signature 11/16/12 Date