



Deep Sea to Coast Connectivity in the Eastern Gulf of Mexico

Data Management Plan

Submitted to the
Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC)

by
Florida State University
Tallahassee, FL

June 6, 2012

Authors:

Jeff Chanton, Professor and Deep-C Data Manager
FSU Department of Earth Ocean and Atmospheric Sciences

Shawn Smith, Research Associate and Data Center Manager
FSU Center for Ocean-Atmospheric Prediction Studies (COAPS)

Tracy Ippolito, Deep-C Coordinator
FSU Center for Ocean-Atmospheric Prediction Studies (COAPS)

The Deep-C (Deep Sea to Coast Connectivity in the Eastern Gulf of Mexico) Consortium is a long-term, interdisciplinary study of deep sea to coast connectivity in the northeastern Gulf of Mexico. The study is investigating the environmental consequences of petroleum hydrocarbon release in the deep Gulf on living marine resources and ecosystem health. Deep-C will examine the geomorphologic, hydrologic, and biogeochemical settings that influence the distribution and fate of the oil and dispersants released during the Deepwater Horizon accident, and use the resulting data for model studies that support improved responses to possible future incidents.

The Deep-C project includes highly integrated research that requires a similarly integrated data management plan to ensure that project execution and data interpretation follow smoothly from the research tasks. Open communication and data sharing are critical to the success of this research. Therefore our data management goals are to ensure the fidelity and accessibility of the Deep-C data, minimize the amount of time research personnel need to spend on data management activities while achieving high quality data and metadata, and ensure that the data and metadata can be located and used by project personnel (initially) and the broader scientific community. Activities that collectively comprise the data management component of the Deep-C Consortium will evolve over time, relying on expertise and tools already developed by the NOAA National Coastal Data Development Center (NCDDC) and by the Florida State University Center for Ocean-Atmospheric Prediction Studies (COAPS).

Our first data management initiative has been to define data collection and distribution requirements, identify common goals with other GRI consortia, ensure that site characterization data are maintained, and resolve any critical knowledge gaps. Data from the Deep-C Consortium will include a combination of empirical data (both observational and experimental), instrumental data (e.g., remotely-sensed, acoustic, and electronic recordings), and model outputs. Deep-C investigators will conform to current NOAA and NSF policies on the dissemination and sharing of research results and adhere to the following best practices:

- (1) Prepare and promptly submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from work.
- (2) Share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work.
- (3) Share software created under the grant or otherwise make them or their products widely available and usable.
- (4) Make results, data and collections available to the scientific community, while observing the principal legal rights to intellectual property.

This preliminary data plan outlines those aspects of the Deep-C data management structure and methods that are known at the time of this reports publication. The Deep-C data center staff anticipates the plan to evolve rapidly in the first few years of the project. Continual evolution is expected as data and metadata collection methods are refined.

I. Deep-C Data Management Points of Contact

The Principal Investigator of the Deep-C project is Eric Chassignet. Dr. Chassignet has more than 25 years experience in research and development, the last 10 of which was spent in formulating, organizing, and managing large, interdisciplinary research programs.

Our Consortium has designated a central Data Manager (Jeff Chanton) who will be in makes decisions regarding overall data management. The data manager will report to the director and works with the five primary research group leaders to ensure that all data generated by Consortium conform to the NCDDC and GRI standards. Our first initiative has been establishment of a data management team consisting of the Data Manager and key Consortium personnel who can provide input and expertise as we work to define data collection and distribution requirements, identify common goals with other GRI consortia, ensure that site characterization data are maintained, and resolve any critical knowledge gaps.

The Deep-C data center will be housed at COAPS with Mr. Shawn Smith directing the day-to-day activities. Mr. Smith has more than 15 years experience managing marine atmospheric and oceanic data through project such as the World Ocean Circulation Experiment, the Shipboard Automated Meteorological and Oceanographic System initiative, and the NSF Rolling Deck to Repository project. Throughout these projects, the COAPS data center has been very successful in developing automated systems for data quality evaluation and dissemination. Mr. Smith will help identify personnel needed to develop an integrated data distribution system for Deep-C and will act as a liaison between Deep-C, the U.S. national data center, and other national marine data management projects.

Deep-C Data Management Team

Jeff Chanton, Professor and Deep-C Data Manager

FSU Department of Earth Ocean and Atmospheric Sciences
117 N Woodward Avenue, Tallahassee, FL 32306-4320
(850) 644-7493
jchanton@fsu.edu

Shawn Smith, Research Associate and Data Center Manager

FSU Center for Ocean-Atmospheric Prediction Studies (COAPS)
200 RM Johnson Bldg., Rm. 236, Tallahassee, FL 32306-2840
(850) 644-6918
smith@coaps.fsu.edu

Tracy Ippolito, Deep-C Coordinator

FSU Center for Ocean-Atmospheric Prediction Studies (COAPS)
200 RM Johnson Bldg., Rm. 236, Tallahassee, FL 32306-2840
(850) 644-6918
tracy@deep-c.org

Diana Villa

FSU Department of Earth Ocean and Atmospheric Sciences
117 N Woodward Avenue, Tallahassee, FL 32306-4320
(850) 644-7493
dianavilla@hotmail.com

The Deep-C data repository will be housed on COAPS servers (currently 150 Tbs, soon to be expended to 1Pb), which will be available to all partners via the Deep-C web portal. All data and model outputs derived by Deep-C researchers will have a copy housed at the FSU data center. These data will be made

available to all partners via the Deep-C web portal. The only exception will be physical samples (e.g., oil, cores, biological samples, etc.), which the Deep-C data center will not accept. The curatorship of physical samples is the responsibility of the individual Deep-C investigators. Data will be transferred to the appropriate national data center (NDC; see section 6) once they are validated and necessary submission metadata have been collected/generated. Provision of the data to the appropriate NDC will release the Deep-C data center from all aspects of long-term (>10 year) data preservation. Before public release, the central Data Manager will verify the readability and validity of the data in close collaboration with the scientists who generated the data. As possible, the public will be provided access to the data via the NDCs.

II. Types of Data

The Consortium's planned scope of work is comprised of five primary research areas: Geomorphology and Habitat Classification, Physical Oceanography, Geochemistry, Ecology, and Modeling. With the exception of the modeling group, each research area will create experimental measurements (many quantitative, but some qualitative), observational data, physical samples, and analytical (lab) measurements. A vast range of instrumentation and analytical tools will be employed. This section provides an overview of the types of observations that will be made within each task area.

Benthic Imaging and Geomorphology Data

Data will be collected along ~10 km scale regions oriented to survey bathymetrically defined features of interest, including erosional or depositional areas, hardgrounds, and natural hydrocarbon seeps. Coverage for geophysical survey will be broad area (polygonal), while benthic imaging data will be situated along linear tracks. The geomorphology team will use imaging systems deployed at the sea floor and from surface ships, aircraft, and satellites. In addition, seafloor mapping and subbottom imagery will be collected. Basic outputs will include bathymetry, sidescan backscatter, seismic profiles, and interpretive maps. Much of the data will be collected during Deep-C cruises using a swath sonar system, CTDs, and a subbottom seismic profiling system.

Processed geophysical survey data will include spatial data as raster geotifs (bathymetry, backscatter mosaics) shapefiles (interpretative products, bottom type maps, navigation), tabular (navigation, acoustic bottom classes), grid files (bathymetry). Map products will be archived as geotif imagery for rapid sharing with GIS systems. Subbottom imagery will be available as annotated jpg files along with matching navigation data. Geologic cross sections will be distributed in Adobe Illustrator EPS format and jpg images. Benthic imaging data will comprise continuous video records, recorded on digital tape and routine (20s) still image files stored as high-resolution jpeg format with embedded geo-reference. Instrument records will include output from a CTD comprising minimally depth, temperature, salinity, turbidity.

Historical data will be used by the geomorphology team. Data types include previous bottom topography surveys (e.g., Okeanos Explorer data) and satellite imagery of surface oil plumes dating back to 1990 (including the Deepwater Horizon spill).

Physical Oceanography Data

Data from the physical oceanography team will be a combination of observations collected during Deep-C cruises and mooring deployments and historical data from the Gulf of Mexico.

Observational data from Deep-C will include CTD surveys, bottom drifters and EM Apex floats, float trajectories, velocity estimations from current meters, and temperature and salinity continuous records from the mooring microcats. CTD data will be processed according to standard procedures and calibrations. Float trajectories will be processed using the ARTOA software. Trajectories will be provided with uncertainties calculated from different parameters (number of sources heard, float position compared to the source array configuration). Instruments on moorings will include ADCPs, current meters, temperature, salinity, and pressure recorders. All mooring data will undergo quality control prior to submission to the Deep-C data center using time series processing developed by SAIC.

Data retrieved via satellite (RAFOS floats, EM Apex, bottom drifters) are downloaded on a daily basis. Other CTD data and current meter data are retrieved directly from the instruments. Hourly, daily, weekly and monthly archived snapshots (ZFS system) will be stored on a regular basis on the COAPS server and made available to GRI partners via the FSU OpeNDAP server.

As an example, the first Deep-C physical oceanography cruise on the RV Pelican, cruise ID PE12-26, included the deployment of six moorings and 10 RAFOS floats. CTDs were conducted at a number of stations. Although some data will be collected in near-real time from these instruments, the bulk will not be obtained until the sensors are retrieved on a future cruise.

One set of historical observations will include data collected by aircraft. These data consist of atmospheric and oceanic profiles of temperature, currents, salinity, winds, and humidity. Data have been collected by the Navy and MMS/BOEM and will be distributed by the Deep-C data center once approval has been received from the principal investigator. Some of these data may be proprietary.

Additional historical data will be collected from moored buoys (e.g., NDBC, COMPS, etc.) in the Gulf of Mexico and the physical oceanographic databases at NODC or NRL. Satellite imagery and altimetry data will be used from public sources. Deep-C will provide an inventory of the data sources being used with data access links when appropriate (it is beyond the scope of the Deep-C data center to house and serve these historical products).

Geochemistry and Hydrocarbon Data

Geochemistry and hydrocarbon data will be collected from sediment cores and from tar samples collected on the shore. Cores will be collected using multicore and boxcore systems during Deep-C research cruises. Core samples will undergo a range of analyses that will create qualitative and quantitative data. Some cores will be split and photographed. Detailed core logs and descriptions will be submitted for each core. Other cores will undergo laboratory analysis. Measurements will include bulk density (g cm^{-3}), short-lived radioisotopes age dating (^{210}Pb , ^{234}Th , ^{137}Cs , ^7Be), mass accumulation rate ($\text{g cm}^{-2} \text{yr}^{-1}$), quantitative sediment texture (% Gravel, % Sand, % Silt, % Clay), and sediment composition (% Carbonate, % Total Organic Matter). Microscope visual descriptions and photography will also be conducted in the lab.

Hydrocarbon samples will be documented by sample location, collection information, and methods of sample preparation. Individual samples will undergo chromatographic and mass spectral analysis in the lab. Data output will include detailed compositional information (class, type, and carbon number) of extractable organic species, as well as classification as acidic, basic, or nonpolar. The hydrocarbon team will collect historical data on biomarkers and abiotic/biotic modification of crude oil from the literature. As possible, this historical data will be distributed by the Deep-C data center for use by other GRI consortia.

Water column data to be collected include CTD profiles, methane concentration and dissolved inorganic carbon as a function of depth. Isotope data will be collected on sedimentary organic matter and dissolved inorganic carbon.

Geochemistry data will consist of relatively small data sets manageable in spreadsheet format such as Excel. These data sets will be organized by the PI's and submitted to through the web portal with the appropriate meta-data for each data set. We will share the primary data with other researchers within a reasonable time or after mutual agreement or through collaboration. Data will be prepared as soon as possible for publication. Our results and data will be made available to the scientific community, while observing the rights to intellectual property. Data that are shared can be redistributed under the provision that the contact information of the person that generated the data, the metadata and the quality control information is associated with the data. After quality control, the data will be submitted to the Deep-C data center and forwarded to the appropriate national archive.

Ecological Data

Ecological data collection will include observations made in the water column and ocean bottom sediments along with samples taken from a range of micro to macro-organisms. Many of the observations will be taken along the DeSoto Canyon and the Florida Panhandle Bight Shelf (FPBS).

Water column observations will include a range of physical oceanographic observations to support ecological research. CTDs will provide temperature, salinity, dissolved oxygen, chlorophyll, CDOM, fluorescence, and turbidity measurements. PAR, UVA, and UVB will be sampled during the profiles. Additional water column measurements include DOC, POC, and TN. Some water samples will be analyzed in the field, while other samples will be captured for shore-side laboratory analysis. Water analysis will determine total and dissolved nutrient contents. Further biological oceanography samples within the water column will include the following:

- Microzooplankton cell counts fixed liter samples surface & Chl max
- Microzooplankton diversity 0.010 mm net plankton vertical tow
- Bacterial cell counts
- Bacterial production by 3H leucine incorporation
- Primary production IR curves with 14C bicarbonate
- >0.2um plankton for DNA extraction surface, bottom and Chl max for microbial ecology/community structure of prokaryotes and eukaryotes: 16-18s rDNA sequences

Additional bottom water samples will be collected using Niskin bottles to study the composition of microorganisms in the water. These water samples will be filtered aboard ship for molecular biology or particulates. Some samples will be analyzed for soluble nutrients and other quantitative measures.

A range of sediment samples will be collected to support ecological studies using multicorer, box core, ShipEx grabs, and/or grab samples. Cores will be sectioned and analyzed to capture a range of information on micro- to macro-organisms residing in the sediment (Table 1). Additional observations from sediment samples include hydrocarbon content, chlorophyll a, POC, TN, DNA extractions for microbial ecology/community structure, PAH content, and oxygen consumption rate.

Table 1: Types of information anticipated from sediment samples.

Research objective	Data or Analyses Planned
Microorganism	Data will include the abundance, distribution, activity, and community composition. Abundance will be determined using molecular techniques or microscopy. Community composition will be determined by sequencing. Activity will be addressed with rate measurements in whole sediment samples that are archived cold for later determinations either onboard the ship or back at the lab.
Microorganism	Benthic chlorophyll and DNA extractions for microbial ecology/community structure, composite of surface sediment for C&N content/ratio, PAH content
Macroorganism	Multicores will be subsampled at 3 depth intervals and each preserved whole in formalin. Preserved samples will be sorted and identified in the laboratory. Counts of each species will be captured in a species by sample matrix and analyzed in PRIMER for multivariate statistics. Samples may also be taken for stable isotopes as available. Representative species will be taken for preserved specimens and provided to the stable isotope group to be run for isotopes. Results will be analyzed to construct a benthic food web, to tie into the larger food web model.

Plankton studies will include net plankton, filtered plankton, and sediment, SEM stubs, LM slides, digital images, cell counts. Net plankton samples will be fixed with Lugol's iodine in the field. A second portion will be cleaned with nitric acid, with small amounts mounted for light microscopy slides using Naphrax and mounted on SEM stubs for scanning electron microscope. Filtered sets, air dried in the field, will include a portion mounted on SEM stubs for scanning electron microscope and a portion mounted for light microscopy. Sediment samples will be fixed with glutaraldehyde in the lab. Portions separated and mounted for (1) light microscopy and (2) on SEM stubs. Another portion will be cleaned with nitric acid and mounted for (1) light microscopy and (2) on SEM stubs for scanning electron microscopy. Light microscope slides and SEM stubs will be used to generate images, diversity measures and cell counts. Some samples will be saved for archival purposes by the PI institution.

Using standardized fisheries practices, longline and trap surveys will be conducted to collect data concerning the distribution, relative abundance and community structure of macrofauna between 50 and 2000 meters deep. Demersal fishes are the primary focus, but we will also collect data on mobile invertebrates. Biological data will be collected on every specimen. In addition, we will collect physical samples from specimens captured for a variety of analyses including toxicology, stable isotopes, heavy metals, life histories, trophic ecology, phylogenetics and functional morphology. Sediment samples and thermocline profiles will be collected at each station.

Individual physical samples (slides, water samples, sections, fish, etc.) will be held by individual investigators in Deep-C for as long as they are required for their research. Sharing of samples within Deep-C

and the other GRI will be encouraged, but the Deep-C data center will not accept or store any physical samples. All ecological data, analyses, or products submitted to the Deep-C data center will be housed at the center until they are cleared for submission to the appropriate national archives.

A range of historical data will be compiled for the ecological research community. Historical hydrographic data will be needed for the Deep-C study region. Plankton studies from past collections by the Deep-C researchers and the literature will be used. Macrofauna information collected previously by the USGS will also be considered along with existing historical data from past fisheries surveys. Locating additional habitat data for the Deep-C study region will aid the Deep-C research and is a suggested task for the GRIIDC.

Modeling Data

The Deep-C Consortium will be generating numerical model output from a number of model simulations. Models to be used include HYCOM, FVCOM, trajectory models, and an oil-fate model. Simulations will include fields of temperature, salinity, velocity, sea surface height, mixed layer depth, and particle trajectories. Model simulations will focus on key study regions (e.g., De Soto Canyon, the Mississippi river plume), while many will also encompass a wider region of the Gulf of Mexico.

The trajectory modeling will take advantage of physical oceanographic data collected as part of Deep-C. The team will use observed trajectories from drifting buoys and also plans to deploy a sail buoy.

The oil fate model relies on Lagrangian particles to simulate the particle advection by the ocean currents. Output will consist of particle positions and oil concentration. Maps will include four dimensional data covering the Gulf of Mexico.

COAPS has considerable expertise disseminating large data sets (i.e. larger than 100 Tbs) to the scientific community. For instance, COAPS has provided such data for the last 5 years to the global and Gulf of Mexico HYCOM prediction systems using four primary component technologies for its data serving: netCDF-CF, THREDDS/ OPeNDAP, and a Live Access Server (LAS). All distribution technologies used are freely available (open source software) and are actively supported by Unidata, OPeNDAP, and NOAA-PMEL, respectively. These components are also used by the Global Ocean Data Assimilation Experiment (GODAE) and more recently by the US Integrated Ocean Observing System as the appropriate standards package for sharing gridded datasets.

The longevity of individual model output on the Deep-C data servers will be determined by the modeling task team. As there is no national data center for ocean model output, Deep-C will maintain access to individual model output for as long as the Deep-C modeling team deems appropriate.

The model simulations developed at Deep-C will depend on a variety of historical data. Individual modeling groups will use climatology fields that are either prepared by the individual groups or accessed from other sources (e.g., atmospheric reanalysis data from NOAA NOMDADS, NRL forcing fields, NOAA NAM, etc). New sources of river inflow data are needed and other inputs from numerical weather prediction models will be utilized. Deep-C will provide an inventory of the data sources being used with data access links when appropriate (it is beyond the scope of the Deep-C data center to house and serve these historical products).

III. Data and Metadata Standards

Metadata are required to trace data origins, determine the types of observational or experimental approaches and processes used, and to validate data. The source of most metadata is the scientists making observations, running models, etc. While metadata collection/generation is often labor intensive, the Deep-C data center will work with the individual investigators to identify the critical metadata for their specific research. We will endeavor to reduce the workload on the scientists by applying forms or other tools to aid the metadata collection/generation. When needed, Deep-C data center personnel will collate and convert metadata into formats/standards desired by the NDCs prior to submission to the archive centers.

Deep-C has developed a cruise workbook (Diana Villa) that will capture a range of metadata for each Deep-C research cruise. The workbook has been developed in Excel and is designed both as a cruise planning tool and a mechanism to capture sampling events during the cruise. The workbook captures a detailed cruise plan with the location of each sampling site; basic metadata for the cruise (vessel, cruise ID, purpose, chief scientist, etc); the sampling events at each site (e.g., multicore, CTD, etc.); and the personnel on the cruise. The workbook maintains a detailed list (controlled vocabulary) of the sites visited by Deep-C personnel. A site can be a point, line, or region to provide flexibility for the activities conducted on Deep-C cruises. Additional controlled vocabularies are used for event types (tasks), vessels, projects, ports, and cruise types. The tool is currently in a beta level and has been tested on the first few Deep-C cruises. The Deep-C data center is working with the workbook developer to enhance the controlled vocabularies, taking advantage of existing vocabularies from BCO-DMO and the Rolling Deck to Repository projects. Workbook contents will be captured at the end of each cruise and imported into a relational database. See *Appendix B* for a Quick Reference guide to the Deep-C cruise workbook.

The Deep-C data center anticipates receiving a range of file formats for both data and metadata (Table 2). Much of the data will be captured in Excel files and some groups will be using databases. The inventory in Table 2 is incomplete and only includes information known when this plan was developed. In some cases, the Deep-C researchers are unfamiliar with metadata formats, so data center personnel will work with the GRIIDC and the scientists to develop mutually beneficial solutions. The data and metadata file formats and standards will evolve throughout the Deep-C project. The management of the file contents and metadata are outlined in Section IV.

Table 2: Anticipated data and metadata file formats for each Deep-C task area. Also noted are standards suggested by scientists within each task area.

Task Area	Known Data File Format	Known Metadata File Formats	Known Metadata Standards
Geomorphology	Unknown	Unknown	<ul style="list-style-type: none"> Unknown
Physical Oceanography	Images and data (NetCDF)	NetCDF	<ul style="list-style-type: none"> Protocols from the IOOS RAs where possible. Some instruments provide standard output (CTD)
Geochemistry	Images (jpegs, PDFs), Excel and custom software	Excel, text	<ul style="list-style-type: none"> Unidata COARDS and CF-1
Ecology	Images (.jpg, .tif) and Excel, MS Access	Excel or Word	<ul style="list-style-type: none"> Seabird CTD output files automatically add metadata. Accepted standard EPA protocols for analytical procedures (e.g. EPA method 3541 for hydrocarbon analysis) Minimum information about a marker gene sequence (MIMARCKS) specifications as defined by the Genomic Standards Consortium.
Modeling	NetCDF, binary, GRIB, Ascii text.	NetCDF, Ascii text	<ul style="list-style-type: none"> HYCOM file standards

The range of metadata required to make the collected information useful to scientists, data centers, and the public varies greatly within Deep-C. Determining the metadata necessary to scientifically describe the observations, analytical samples, and model output will be driven by the expert researchers within Deep-C. Table 3 provides a subset of the metadata that scientists in each task area listed as useful for their research activities. Again, this list will evolve as the project develops. The Deep-C data center will work with the research task teams to develop metadata protocols suitable for their research. Whenever possible, we will utilize metadata standards (see Table 1); however, many of the researchers are not familiar with standards, ontologies, or vocabularies. We will work with the GRIIDC and the task teams to identify standards and implement them in their research activities. We anticipate that some cross-walking of individual science party vocabularies to national standard vocabularies will be required and this will be accomplished by the Deep-C data center within our data/metadata tracking database.

Table 3: Preliminary list of metadata that science teams have identified as important for the scientific application of their data and analyses.

Task Area	Metadata
Geomorphology	Vessel, cruise number date, time (UTC); sub-region (e.g. transect) and associated repeat sampling sites; default spatial parameters; instrument models, serial numbers, and calibration; camera types, lenses, scaling lasers, storage media; data-storage media, serial numbers, and archive files; and, Cruise participants.
Physical Oceanography	Time (UTC), latitude, longitude, depth, SI units
Geochemistry	Cruise number/name, station number, date, time, time of sampling, sample location, collection information, sample preparation, chromatographic and mass spectral conditions
Ecology	Cruise number/name, station number, date, time, time of sampling, either as an instance or incubation interval, longitude/latitude, transect, region, depth of sample, collection type, net characteristics, filter characteristics, collection gear, parameter measured, method used, unit, person doing the sampling/analysis, microscope type/magnification, counting protocol, sample treatment (frozen, stored cold, filtered or not), environment (sediment, water, etc.), sequencing (target gene, sequencing method), experimental treatment (in the case of amendments to incubations), CTD data; Oxygen; sediment grain size; sediment OC content; overlying productivity; POC flux to seafloor; Isotope values for all potential end members of the food web mixing model
Modeling	Depths, dates, latitude, longitude, SI units, quality controlled or not, instruments types, and experimental set-up

Several challenges and knowledge gaps have been identified. Metadata standards for many data types are unknown to the Deep-C community, so solutions will need to be identified. For example, several modeling groups will be disseminating output in a variety of formats. The NetCDF CF conventions would be one standard to adopt, but this may not suit the needs of all modeling project for Deep-C. The Deep-C data center will collaborate with GRIIDC and other data project (e.g., NSF R2R, OceanSites, BCO-DMO, etc. to locate suitable standards for the wide range of data and products being created by Deep-C researchers. Another critical need for Deep-C is to identify a standard protocol for the handling and archival of digital imagery. Digital imagery will be produced by all of the task teams within Deep-C, including video files from some groups (modeling), and metadata standards are needed to ensure proper image file tracking and provenance. Finally, several groups have noted that the plan to provide metadata in MS Word files or cruise reports. The Deep-C data center will work with these groups to try to get their metadata into a digital format that is easier to parse electronically (e.g., Excel, csv text, etc.).

IV. Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

Data files, sample analyses, and model output will be housed at COAPS on a petabyte data storage system. All data sets will be tracked through using structured query language (SQL) databases and many of the smaller data sets will be stored directly in the database. Larger model and satellite data files will be stored in netCDF files outside of the database. Using a combination of SQL databases and netCDF files will facilitate distribution of the data using a range of geospatial access tools.

All Deep-C data will be made available through a series of web services. One tool we will employ is the live access server (LAS), a configurable scientific data "product" server (see Figure 1) that provides a user-friendly interface to browse gridded ocean/atmosphere data and model outputs. With nothing more than a Web browser and an Internet connection, a user can obtain data and perform simple analyses, producing plots, images, and formatted files generated on the fly from custom subsets of variables. Using the LAS comparison mode allows users to differentiate (with automated re-gridding) variables, overlay them graphically, and view them as side-by-side plots.

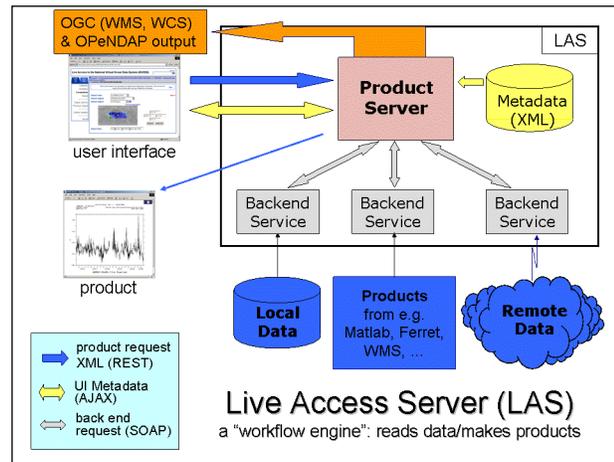


Figure 1. LAS users can view visualizations, obtain subsets and perform simple analyses without special software tools.

We will also establish a geospatial web service for a range of the Deep-C data. COAPS currently operates a map server which has been used to develop an interactive online atlas for the Gulf of Mexico. The capabilities of this map server will be augmented to allow display of Deep-C data from cruises, onshore sampling, moorings, floats, gliders, etc. The map server will be connected to our SQL database and a THREDDS server for the netCDF files to provide the user a range of options to view and download Deep-C data and model output. The details of the database and map server structure will be developed over the coming months.

Deep-C is collecting a wide range of data types and some will be subject to embargos. Overall, the Deep-C investigators will conform to current NOAA and NSF policies on the dissemination and sharing of research results and adhere to the following best practices:

1. Prepare and promptly submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from work.
2. Share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work.
3. Share software created under the grant or otherwise make them or their products widely available and usable.
4. Make results, data and collections available to the scientific community, while observing the principal legal rights to intellectual property.

Permission to access the Deep-C data will occur at four levels: (1) the investigator collecting/creating the observation or model output, (2) within the Deep-C consortium, (3) among the GRI consortia, and (4) the general public. As we proceed with development of the database to track individual data sets and model output, mechanisms will be put in place to provide release authorization for each of these levels. Each Deep-C investigator will be required to set a release schedule for different data types and the policy will be refined and approved by the Deep-C steering panel. We anticipate that many data sets will be made available at all levels soon after data collection, but some data may be held to allow individual investigators time for initial publication of their results.

Basic discovery metadata for all data sets and model output will be made available even for those data sets that are under an embargo. The policy of Deep-C is for the provision of discovery metadata along with an indication of the length of the data set embargo to the Deep-C data center by the responsible PI no later than 60 days after sample collection or instrument deployment (e.g., info on moorings, drifters, etc). The metadata request will also require the PI to provide a timeline for analysis of the samples or observations with a deadline for submission of the resulting data or products to the data center. All metadata information will be stored in an SQL database and will be made available via web services to all interested parties as soon as possible after receipt of the metadata from the individual investigators. The data center will also establish a data submission protocol that will include a requirement to capture essential metadata for each data set submitted to the center. The Deep-C data center staff will work with GRIIDC personnel to exchange any metadata in the formats desired by GRIIDC.

Deep-C has established a policy to protect intellectual property and publication rights. The full policy document is attached in *Appendix A*. In addition, Deep-C researchers will ensure that all parties engaged in research involving the use of live vertebrate animals have an approved Animal Welfare Assurance and that the activity has valid Institutional Animal Care and use Committee approval (e.g., <http://www.research.fsu.edu/acuc/>).

Deep-C will adopt the policy on personally identifying information (PII) established by the NSF Rolling Deck to Repository project. Research vessel cruises (and any other field research) data and documentation submitted to the Deep-C data center will be deposited in appropriate National Data Centers for permanent archiving and dissemination. As such, they must comply with federal guidelines including the U.S. Department of Commerce (DoC) policy on Electronic Transmission of Personally Identifiable Information (PII) (ref. U.S. Office of Management and Budget, Memorandum M-07-16, "Safeguarding Against and Responding to the Breach of Personally Identifiable Information").

Per DoC policy as published by the Office of the Chief Information Officer (OCIO), PII is defined as "information which can be used to distinguish or trace an individual's identity, .. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual." PII is further classified as "Sensitive" or "Non-Sensitive".

"Sensitive" PII includes:

- Social Security Numbers (including truncated to last 4 digits)
- Place of birth
- Date of birth
- Mother's maiden name
- Biometric information
- Medical information, except brief references to absences from work
- Personal financial information

- Credit card or purchase card account numbers
- Passport numbers
- Potentially sensitive employment information eg. personnel ratings, disciplinary actions, and result of background investigations
- Criminal history
- Any information that may stigmatize or adversely affect an individual

Data and documentation submitted by Deep-C researchers, should not include any PII classified as “Sensitive”. It is the responsibility of the individual researchers to remove any PII prior to submission and any PII located by Deep-C data center staff will be deleted upon receipt.

Note, for reference, the Deep-C data center will accept “Non-Sensitive” PII which includes the following:

- Work, home, and mobile phone numbers
- Work and home addresses
- Work and personal e-mail addresses
- Resumes that do not include an SSN or where the SSN is redacted
- General background information eg. Biographies
- Position descriptions and performance plans without ratings

V. Policies and Provisions for Re-use, Re-distribution

As noted above the Deep-C data center will establish a protocol for scientific embargos for individual data types. Once any embargos are cleared, there will be no restrictions on access to data submitted to the Deep-C data center. Open communication and data sharing is critical to the success of the Deep-C research.

The range of future users of the data and model output from the Deep-C research is unknown. We anticipate interest from researchers in the geosciences, biologic and environmental sciences, and political sciences. The data will be useful to both academic and private sector communities. In addition, the data and results will be of interest to formal and informal educators throughout the Gulf states.

Since all future uses of the data cannot be foreseen, the Deep-C data center will strive to collect the range of metadata required to describe the data long after the conclusion of the project. Submission of the data to the national data centers will preserve the legacy of the project for the foreseeable future.

VI. Plans for Archiving and Preservation of Access

The Deep-C data center will be responsible for maintaining a copy of all data and model output during the project. As data embargos are cleared, the data will be submitted to the appropriate national data center (NDC). The Deep-C data center staff will work with GRIIDC to determine appropriate national archive centers and to establish the necessary submission agreements with each NDC. Although no data submission agreements are currently in place, we anticipate data from the individual task areas will be submitted to the NDCs that follow:

1. Benthic imaging and geomorphology data - National Geophysical Data Center (NGDC)
2. Physical transport data – National Oceanographic Data Center (NODC)

3. Geochemistry and hydrocarbon data – National Coastal Data Development Center (NCDDC), NODC
4. Ecological data – NCDDC, National Center for Biotechnology Information, GenBank, EMBL Sequence Nucleotide Database
5. Model data sets – No national data archive exists for model data

The long term curatorship of Deep-C data will be the responsibility of the NDCs. The Deep-C data center has no plans for the necessary technology refresh and digital media migration to ensure the integrity of any data or model output beyond the end of the Deep-C project. We do anticipate that some data orphans, those data not accepted by any NDC, will exist and the Deep-C data center will endeavor to maintain electronic access to these orphans for as long as resources allow support for necessary infrastructure.

The Deep-C data center anticipates some data and metadata transformations will be necessary. We will construct the necessary metadata records (ISO, etc.) desired by the NDCs prior to submission of data sets. We anticipate close collaboration with the GRIIDC to develop the necessary tools for metadata record development. As for data files, the Deep-C data center will archive them in the form that they are originally received at the data center. We do not anticipate the NDCs requiring any data transformation; however, when the Deep-C data center transforms any data received, the necessary data provenance will be included to ensure that the link to the original file is maintained.

The role of the Deep-C data center is to collect and disseminate the data and model output provided to the center by the project researchers. The data center will not clean or scientifically quality control any data, this is the responsibility of the project research staff. The Deep-C data center will make provisions to track the file versions for any data/model output submitted to the center. This will allow research staff to submit the original (raw, as collected) data and subsequently submit quality controlled, processed, or cleaned versions of the data. Although the original data will be made available on request, the Deep-C data center will provide the quality controlled, processed, or cleaned data to the user community through and data distribution systems. The submission of data to the NDCs will include the original and any subsequent processed data sets, with appropriate file provenance and metadata.

The nature of the research being conducted is anticipated to result in ancillary documentation. These documents may include cruise reports, event logs, data processing software, model code, etc. When these are submitted to the Deep-C data center they will be required to include metadata necessary to trace them to a specific Deep-C funded activity. If these documents are desired by NDCs, they will be submitted along with the appropriate data. If not, they will be housed and distributed by the Deep-C data center.

If desired by the Deep-C steering committee, the Deep-C data center will provide an online software library that will be accessible to the user community. Note that the Deep-C will not distribute any program codes that are not considered open source. Any proprietary codes should not be submitted to the Deep-C data center. No effort will be made by the Deep-C staff to document or provide user support for these codes. This will be the responsibility of the code authors.

Appendix A:
Intellectual Property and Publications Policy

Appendix B:
Deep-C Cruise Workbook
Quick Reference