

Coastal Biophysical Inventory
Report to the California Department of Fish and Game

Joseph Kinyon, Pacific Coast Science and Learning Center
12/15/2008

Introduction:

This report is a summary of the Coastal Biophysical Inventory database. The report includes the digital files and datasets that accompany this summary. The report includes the following items:

1. Summary
2. Protocol
3. Coastal Biophysical Inventory
 - a. Observations of coastline in a Microsoft Access 2003 custom database
 - b. Digital Field Photos
 - c. Digitized Field Data Sheets
 - d. CBI GIS
 - i. ArcGIS 9.2 Personal Geodatabase
 1. Segment Locations
 2. Reference Layers
 3. Raster Catalogs of DRG
 4. Raster Catalogs of RNC
 5. Raster Catalogs of Aerial Imagery
 - ii. Flat file storage of GIS raster and vector data
 - iii. GPS data dictionary
4. Characterization of the Data Set and Organization themes.
5. Instructions for installation of software and Demo
6. Data Entry Instructions
7. Pointers for relating observation database to spatial database in an ArcGIS 9.2 map document.

Summary:

In 2002 Point Reyes National Seashore (PORE) and Golden Gate National Recreation Area (GOGA), formed a cooperative agreement with University of California, Davis, to conduct a coastal biophysical inventory project. This project included substrate, marine algae, and invertebrate surveys conducted in the intertidal. The field protocol for this project was modeled after similar surveys conducted from 1995-98 by Glacier Bay National Park in Alaska. In 2006, a cooperative agreement with the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) group at the University of California, Santa Cruz was made to extend and complete this Coastal Biophysical Inventory.

Coastal segments were delineated based on substrate and slope, and additional physical and biological data were then collected, including GPS boundary coordinates, the location of representative biological transects, and various physical and biological observations of the coastline. All field work was conducted during low tides within each segment.

Total coastline observed is approximately 161 kilometers. The inventory was restricted to coastline managed by or in partnership with the Point Reyes National Seashore and the Golden Gate National Recreation Area. There are notable exceptions to this and the locations of coastline excluded are Drakes Estero, Tomales Bay, Abbotts Lagoon, Bolinas Lagoon and Rodeo Lagoon.

These data were compiled into the Coastal Biophysical Inventory Database (Microsoft Access). This user-friendly database allows researchers, park managers, and the public to “virtually walk” the coast via maps and ground photos, and to view both biotic and abiotic information about the parks’ coastal resources. These data will be used for: 1) natural resource damage assessments; 2) biological and physical resource inventories; 3) baseline data for monitoring of selected species and habitat types; 4) identification of rare habitats or species; and 5) pilot data for studies on coastal patterns of biodiversity.

Protocol:

Please see the attached document *Coastal Biophysical Inventory (CBI) and Mapping Protocol* or the enclosed file “CBI_Protocol_Draft_Version_A_121208.doc “.

Coastal Biophysical Inventory:

The Coastal Biophysical Inventory is an electronic data set of observations and GIS data. It is approximately 9 gigabytes in size. All components are located on the CBI_PCSCD DVD in the folder “Coastal Biophysical Inventory.”

Observation data, photographs and scans of original datasheets are best accessed using the Microsoft Access 2003 database “CBI_OBSERVATIONS_PCSCD_NPS.mdb”

located in the CBI_Dataset directory found in the Coastal_Biophysical_Inventory directory.

GIS data is stored in the CBI_GIS directory and is designed to be accessed or manipulated using ArcGIS 9.2 desktop tools in a Windows XP service pack 3 environment.

An enclosed Personal Geodatabase “CBI_GIS_PCSLC_NPS.mdb” makes access to the GIS files easier. Several example Map projects linking to the data are enclosed as MXD documents that make use of this personal geodatabase.

It is likely that database connections and file linkages will need to be repaired once first installed on a new machine.

In order to use the Observation Database, the custom CBI database software must be installed and a minimum of base software must be installed.

Operation of Essential Features

The database requires a base level of hardware and software to operate. It requires the user to install supporting libraries and objects used by the software in its calculations and operation. After it is installed the essential features of this data base is the facilitation of the methodical organization of data collected in the field through its databases and file structure.

Installation of the user libraries is accomplished through a custom installer package. It is an automatic installation of the components required by the software such as Microsoft Data Access libraries and ESRI MapObjects ActiveX objects. This installer saves the user from the tedious activity of installing or registering each programming library or object individually. While the installation instructions are detailed to take care of installation anomalies, the installation is a rather simple set of steps including copying the database to its location on the computer, running the setup application, interacting with the application set up dialogue until it is finished, and running the database.

Tabular data is entered and maintained through the Data Entry page of the User interface. Because the Data Entry page is an Access Database, operation of the database and manipulation of the forms, tables, queries, reports and modules should be familiar users of Access databases.

Management of the imagery data is almost exclusively accomplished through the built in cutting, pasting, and renaming tools of the Windows XP GUI. The user can easily copy digital files from the camera or media using Windows Explorer.

Management of the GIS dataset in the Personal Geodatabase and the related folder structure is done through ArcCatalog and ArcGIS. While it is an important part of the deliverables for this project and integral to the final deployment of the Coastal Biophysical Inventory, it is beyond the scope of this report to describe the operation of the GIS beyond the data management description and workflow listed in the protocol

Data Entry Instructions:

The documents DataEntry.pdf or DataEntry.ppt give an introduction to data entry for those familiar with Microsoft Access. It is not the intent of this document to train the user to be a competent user of Access, only as a roadmap to users familiar with this database software.

Characterization of the Data Set and Organization themes.

This dataset is an adaptation of previous protocol and the project has had two managers and spans approximately 8 years of planning and implementation. For this reason there are idiosyncrasies that were inherited and introduced when making improvements to the written protocol.

The current database takes its design from the revised protocol. Besides explicit clarity for the steps to characterize a segment, the most significant change is the adoption of the robust Taxonomy practices of the PISCO intertidal team at UCSC and the adaptation of the protocol for locations that are inaccessible by foot or beach landing.

The taxonomic changes required a process of lumping and splitting the names of the previous inventory staff to improve and match the new taxonomic expectations. Legacy data is preserved in the database in its original form as a reference. A handy reference file “species_list_comparison_summary070308.xls” also shows the “crosswalking” of the legacy data. Whenever reviewing the data for species observations or assessment of biodiversity is essential to refer to this documentation of the taxonomic changes made to data collected prior to July 1, 2007. Taxonomic data has been manipulated to conform to the new taxonomic standard.

A statistical analysis of the data collected can be found in the file “Exploratory analysis to examine robustness of dataset_v1.3.pdf”

Access database

Rationale to design choices:

This section describes the design of the database that holds the Coastal Biophysical Inventory tabular data.

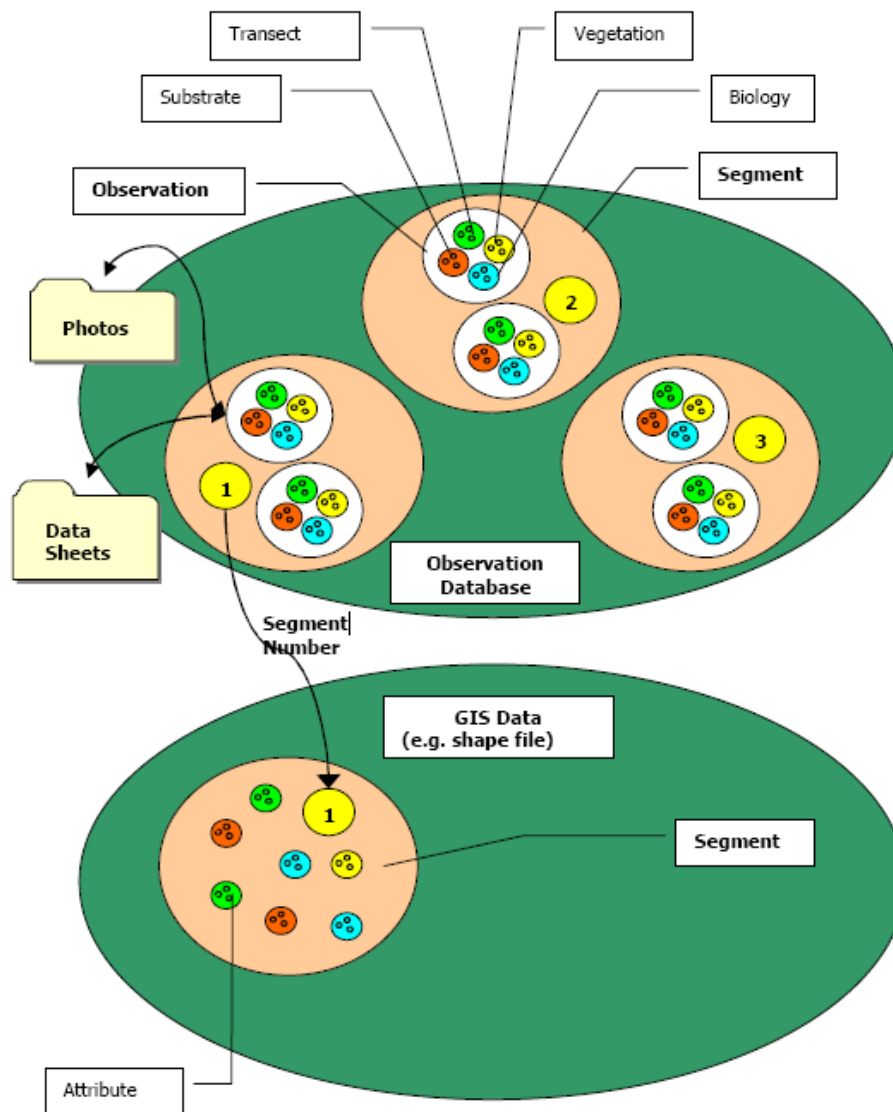
The database design was distilled down to six essential parameters:

1. It must hold the data type created.
2. It must be relational.
3. It must be able link to the external GIS data
4. It must be able to link to the external objects data.
5. It must reinforce the integrity of the data as defined by protocol.
6. It must be replicated at different locations.

1. There are many types of data, text, integers, numbers with double precision, binary raster data as images, paper documents, coordinates, spatial features, etc. While almost all formats can be held within an Access Database, it excels at storing text and integers in tables.
2. Because the essential goal is to observe the coast before damage occurs in order to compare it to observations after damage occurs, the data is at its origin a many to one relationship of before and after observations to a place.
3. Based on the protocol, the organizing structure to the location information of the data is substrate. The combination of different sizes and shapes of substrate define it as a place. Each place becomes the pivot for the database, and the unique number for this place becomes the relational link between the tabular data to other tabular data and GIS data.
4. Because of size limits and performance limits to the Access format database, larger data types such as the GIS and image files need to be stored external to the database, but fixed in location relative to the database so that it can retrieve the data using paths that are relative the database files own location.
5. With large amounts of data being transcribed from field data sheets to the database, there is great opportunity for data entry error. Validating or restricting the data options greatly improves the integrity of the data by excluding the opportunity for data entry error.
6. The database must be able to allow users at two different locations enter data. Because of its size, online hosting was not an option and due to unreliable network connectivity and FTP site integrity the database could not be situated in one spot with multiple concurrent connections. Data entered into one must be able to be inhaled into the other.

Conceptual structure:

The segment, in a sense, holds the observation. However, the segment does not exist until the first observation instantiates the segment by the field biologist classifying the substrate at a location. This act of classifying the extent of a type of substrate bootstraps the segment into existence. From then on, the segment can contain observations of that location and relate to other data. Each observation holds multiple attributes. The result is that the database holds many segments, each segment holds many observations and each observation holds many attributes. It is much like the Russian folk toy of nesting Russian dolls. In fact it is like whole families of nesting dolls fitting into one big doll.



Maintaining Referential Integrity

The nesting is organized by a cascade of linkages. The first and primary table in the database is the Segment Table. It holds an indexed collection of unique identification numbers for each and every segment. This unique number is created in the field when the researcher types a part of the coastline to a particular substrate combination. A data sheet of observations is given the number and GPS coordinates are collected for its upcoast and downcoast limits and given this same number. (Upcoast and Downcoast are terms of topology applied to the Coastal Biophysical Inventory. Upcoast is to one's right while standing on land and facing the sea, downcoast is to one's left.) The table holding the segment information will

not allow duplicate segment numbers, so that if a field observation is accidentally given the same number twice, then it can not be entered until the GPS coordinate's segment identifier and the datasheet's segment identifier are both changed to a new and previously unused number.

The second and most important table is the Observation Table. Each observation is assigned a Global Unique Identifier (GUID) and the Segment Id number to which the observation is about is attached to that GUID in the same observation record. The GUID allows for unique sequential numbers and is the key to linking the related observation data. A GUID is essential for combining replica databases from two different locations into one. If only a sequential unique number is used, then when the databases are combined on a later date, the data fails data validation because the identifiers for the observations are most likely no longer unique (there is an excellent chance both data sets start numbering at 1). In theory, the GUID is based on unique combination of a number of serial numbers unique to the user's computer's network id card, processor, username, etc. This means that no two machines will ever enter the same GUID as any other machine, nor will it ever repeat a GUID on the same machine. It maintains data integrity because the data manager can combine data into a database without conflating data, keeping each unique observation record unique. In the off chance that a duplicate GUID is created, the existing database will reject any GUID that has already been added. This is useful for two reasons. It catches the wildly unlikely chance of duplicate GUIDs. Also, If an "append to table" batch process is accidentally done twice, it rejects any duplicates. The data that relates to the observation in a one to one way, for example one date for observations do not span more than one day according to the protocol, are also entered in this table to keep the database compact in size, simpler in its complexity, reduced to its essential elements in structure. This is also known as normalized.

This Observation Table uses a double primary key. The combination of the unique segment number and the GUID observation number reinforces the ability of the data to make many changes, have multiple updates from replica databases, and reinforces referential integrity of the database during cascading deletes or updates without duplication or mis-relation of data to other tables. Each table that connects to the Observation Table (e.g. Photos, Transect, Biology, Streams, Vegetation, etc.) and is a subset of that observation also has the same double key. The double key is a denormalized trait designed into the database. The database could make use simply of the GUID key linked to the Observation table and related tables, however if one wanted to simply sort the Biology Table by Segment ID they would have to build a more complicated query to join the three tables together through the GUID and the Segment ID. GUIDs are not user friendly to read, they are machine friendly to create. The Segment Id tagged on to each GUID helps a viewer quickly browse related tables visually and see the relationship to the Segment immediately. For example if a Segment number is 02DS04 (the fourth observation is made by Dave Schirokauer in 2002), then the GUID for an observation for that segment would read "BD2F6549-113F-4EA9-94E8-5447ED486EB6". The use of the double key also allows connection of observation tables to external GIS data without lengthy Structured Query Language

(SQL) format sentences that are subject to failure due to syntax or spelling errors in their construction. These were the essential parts of the decision to denormalize parts of the database in order to ease data table browsing and the linking to outside data sets with simpler SQL query sentence structure

Each table that relates to the Observation Tables has the double key described above, but only the Observation table makes that specific double key a double Primary Key. The related tables use the GUID and the main observation field for that data as a double Primary Key. For example, the Biology Table has a GUID field to connect to the observation and a Species Name field for the observation. The Observation table's GUID is a unique number linking the Biology Table. Because there can be multiple observations for the same segment, the Segment key allows duplicates in the Biology Table. However, one would never want to enter 4 Ochre Stars, 4 Crabs and then 2 Ochre Stars for the same observation. It should be entered as 6 Ochre Stars and 4 Crabs. If a data entry person isn't paying attention, they might enter the same crab observation twice in a row after stopping to answer the phone and picking up where they left off. By combining the Species Name and the Observation GUID into a dual Primary Key, the database rejects incorrect duplicate entries, because the combination of these two is set to be a unique pair.

Referential Integrity, cascading deletes and cascading updates can happen in this database because the relationships are built upon these double keys throughout the database. If someone deletes a Segment, then all observations for that segment are deleted. Using the nested doll analogy again, if you throw out the big doll you also discard the many smaller internal dolls. That is a cascading delete. A cascading update occurs if the user updates the Segment Id, then all related observations and the Observations Table's related tables also update. This is unlike a nested doll. If you paint a mustache on a nested doll's biggest doll, all the smaller dolls inside do not get mustachioed too. This is quite helpful for a database manager who needs to correct or update the spelling of something. Change the database in one place, and all the places in the database where it should be updated are automatically corrected. (See the Database Relationship Diagram).

Legacy Data

The remaining database design includes three other types of tables. They are legacy data held in its original format for reference, tables used for holding the results of queries temporarily, and domains in the form of look up tables. Legacy data is important because many choices were made when transforming the data and conforming it to the revised protocol. Some of these revisions were simple, such as splitting the Wave Swell field that read " 2 m 312 deg" into a Swell Meters = 2 and Swell Direction = 312, others required changing a binary "Yes, No" to a string of choices " Yes, No, Not observed, Not Applicable." Maintaining the legacy data allows a data miner to observe the originally entered data.

Tables as programming tools

Some controls on the user interface can not connect to virtual tables held in the computers memory as a record set or array; they are designed to only connect to other objects in the database. An example of this is the drop down combo box used liberally throughout the database interface. Usually the source for the list is obtained from a table that can be updated. Data Entry Person's Name is a good example of this. When a user updates the Data Entry Person's Name table, the combo box is updated. The drop down combo control can also be filled programmatically when it is first created. If a database user creates a record set in memory and fills it with the results of a query (e.g. SELECT all Segments that have Seal Sleeping Areas in their observations and sort the row alphabetically) the combo box can not fill itself with the results of that query because it does not have access to the reference for that memory location in the Visual Basic for Applications language. The solution is to create a table to hold the results of queries like this temporarily so that the combo box can have instant updates.

Domains

Domains are the expert data. Domains help maintain data integrity similar to the referential integrity settings using the Primary Keys described above. All the domain tables in this database are called "Look Up" tables. If I type my own name 150 times during the course of data entry, there is a pretty good chance of making a mistake typing it incorrectly. If there are hundred's of fields on the database with even lengthier entries such as *Anthopleura xanthogrammica* the risk of incorrect spelling becomes an unacceptable risk. In many cases throughout the database data entry user interface, the combo box auto completes and autocorrects the data entry person's typing to match the list of accepted spellings. Limiting the possible options of entry makes data entry easier and maintains a higher level of error free data.

Because human experts will be viewing the data and the fact that there is plenty of expected room in the database for growth, a few aspects of the database were left denormalized. In a few situations the data as left denormalized in its idiosyncratic format because the complexity of converting the legacy data to the new desired format was more effort than it was worth compared to other project demands. Other changes were not made with the "If it isn't broke don't fix it" approach to project management. This was true for the biology data.

Taxonomy and database design

Creating a unique taxonomic identifier was not included in the design of the Biology section of the database. Taxonomic serial numbers relate to the most precise identification one can give to an organism. If one can identify and an organism down to species in the field, a species name is noted and the taxonomic identifier is entered into the database. Occasionally species names are revised; the species is split into two species, lumped into one, or reverts to a name given

previously in time. If the taxonomic id is used, then when the species name is updated on the lookup table, no cascading update is required. In this database the full text of the organism name is given and this additional use of characters repeatedly “bloats” the database and makes it less efficient to search than a database table designed with a unique and indexed serial number. The performance improvement for this database would be negligible if changed, so the decision was made to leave the conceptualization of the taxonomic data as it is with potential to change at a later date in order to relate it with the growing success of All Taxa Biotic Inventories and taxonomic names databases such as the Integrated Taxonomic Information System (ITIS). Currently these databases are inadequate for the complexities of the taxonomy of marine invertebrates and do not list a majority of the organisms we regularly encounter. Our taxonomic names standards now conform to the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) current taxonomic tables provided by the PISCO group at UCSC.

Reference Images

Finally, due to the total size of the digital photo dataset increasing with each and every segment and the resolution doubling for each camera model used through out the history of the project, large binary data (aka. BLOBS) were designed out of the database. The other BLOBs excluded were the PDF format files of every paper field data sheet. A minimum of four photos were taken of each segment and two datasheets for each observation. At an estimated 800 segments, this will be a large dataset. The database was designed to relate to a tree of folders adjacent to and below the database. This was explicitly done so that relative paths to the objects could be stored in the database as file paths in a text format. Programmatically the database can see where it is in the file structure of the computer and then access these related files through their relative paths. This means that JPEG or PDF format files remain stored in their compressed state rather than being enlarged into their uncompressed state in an Access database as a BLOB, making the final entire dataset much smaller.

GIS Database:

The GIS database was much simpler to design, as most of its design was outside of my control. I adopted an ArcGIS 9.2 Personal Geodatabase as the primary GIS database format. Because the high resolution imagery is much greater than a Personal Geodatabase can hold, the raster imagery were linked as raster catalogs in their unmanaged format. Deleting the raster catalog in the Personal Geodatabase does not delete the raster object occupying the storage media. Because relative linking is expected for this database, and portability is also expected, the GIS data was organized like the imagery data. It sits in folders adjacent to or below the Personal Geodatabase. This is so the database can quickly repair its references to linked data stored in the neighboring folders if it is moved to another drive and is accessed from an ArcGIS Map Document. While this solution doesn't eliminate all problems with moving larger linked geodatabases, it

greatly enhances the portability of the whole dataset.

The GIS database did receive one pivotal improvement and this was the creation of a specific piece of data for the database. The original line feature for each segment was created by “snipping” or breaking the existing coastline created by the National Wetlands Inventory mapping project. Because the intertidal zone is a type of wetland, linking it directly to this dataset and coastline features made sense. When reviewing the data it became clear that the use of the NWI line feature had been inconsistent, with some sections of the coast derived from administrative boundaries or USGS coastline features. It was clear that a well documented coastline created from a consistent data set was required. For this process I generated a seamless grid of 10 meter pixel Digital Elevation Model data from the USGS for the coastline. From this I created a shoreline by converting all pixels above 0 meters into a vector feature. The Coastal Biophysical Inventory’s central piece of data is a set of polyline features generated from the shoreline which is created by breaking the line feature at the points closest to the GPS coordinate that marks the upcoast or downcoast section of the segment. Each line feature receives the attribute of the segment id and this is the manifestation of the Coastal Biophysical Inventory in the GIS portion of the database.

Metadata is stored within the Personal Geodatabase and along side the GIS file in the folders. The data metadata is stored using Metadata Standards adopted for the Pacific Coast Science and Learning Center. The metadata is stored in an XML file and complies to the FGDC standards first, National Park Service Standards second and Pacific Coast Science and Learning Center Standards third. These standards were developed as part of the database design.

Relating Observational Data to the GIS

Please refer to the document “Relating tables in ArcGIS 9.doc” for instructions on the most straightforward way to link the data in an ArcMap document using ArcGIS tools.

Installation of Software:

For this software you will need:

1. A monitor running 1280x1024 landscape (This database can also be viewed on
2. 1600x1200 or 1200x1600 resolution monitors)
3. An active high speed wireless connection for viewing Google Map or California
4. Coastal Records Projects photos.
5. A Windows XP machine running SP 2, with the Operating System installed on the C: drive (default installation)
6. Access 2003 (sp3 recommended)
7. Microsoft Excel 2000 or greater.
8. A Zip file compression program
9. Internet Explorer 6 or Firefox 2 browser installed (Google Chrome works too)
10. Microsoft Document and Fax viewer turned on or installed (Usually part of the default XP installation)
11. Adobe Reader 7 or greater.
12. Approximately 12 GB of free space for the files

Database Setup Steps (please note that this setup may require a restart):

1. Copy the Coastal Biophysical Inventory package to the folder of choice (it is designed to run wherever it is moved).
2. Right mouse click on the CBI folder and select Properties at the bottom of the dialog menu.
3. On the Properties menu make sure that the check box for "Read Only" under the attributes section is NOT checked (If this is not done the startup will have an error because the Setup.lst file can not be read.)
4. When the confirm attribute changes dialog comes up, choose the bottom option "Apply Changes to this folder, subfolders and files."
5. Inside the Coastal_Biophysical_Inventory folder there are three folders CBI_SETUP, CBI_Dataset, and DFG Report
6. Run the setup.exe program in CBI_SETUP, you will return to CBI_Dataset later.
7. The setup program will prompt you if it finds any file that you already have that is a more recent version. I would say no to any of these replacements, especially if you are running a MapObjects LT 2.0 design environment.
8. Once the setup is complete, it may prompt you for a restart, please restart your computer.
9. This setup will not create short cuts to the database file, create a shortcut to the database and place it in the Coastal Biophysical Inventory Start Programs Directory
10. For a list of 17 libraries and ActiveX objects that will be installed, please review SETUP.lst file under the CBI_Setup folder using Microsoft Notepad.

Application Notes:

1. In the folder CBI_Dataset is the Access 2003 database
2. CBI_OBSERVATIONS_PCSLC_NPS.mdb
3. Open this file (do not move the database file from this location.)
4. When prompted, accept the macros in this database and select OPEN.
5. This will open the database and set the views to optimize the window.
6. The database will open to the Welcome screen.
7. You can double click on the initial NPS splash screen to close it.

Programmers Notes:

1. Normal Access menus are turned off. If you need to view these, select Close Form from the switchboard on the left side OR hold down the shift key when selecting open at startup.
2. To view the raw code in the VBA programming environment, hold the Alt key and F11 key down with the FLOCK turned on after opening the database in full mode (holding the shift key down during the opening sequence). Please only do this on a copy of the database because if you do not have developer licenses for some of the objects it can corrupt the form.

Orientation to Database:

1. *NOTE: Instructions for navigating the database are not included in this delivery.
2. Database navigation is primarily done through the switchboard column on the left side of the screen.
3. The data entry interface mirrors the hierarchical structure of the data with main forms holding sub-forms. The main form is the Segment form, with ~425 records.
4. Each segment can have an observation record, and each observation has a set of Substrate, Features, Biology and Photo information connected to it.
5. When navigating the Segment records using the buttons at the bottom left, if the observation form is blank inside it, this means the observation has not been entered yet.
6. Light orange colored fields indicate data that is "missing" and is a visual cue that it needs to be updated or entered.
7. You can use the drop down box below the segment number to advance to a specific segment.
8. Currently the database is optimized for data entry, not data review.
9. The page Documentation has links to many supporting documents, including blank data sheets and an older draft of the protocol rewritten with Rani Gaddam of PISCO at UCSC.