# Observation Data Model (ODM) For Rincon Bayou, Nueces Delta

Final Report By:

Paul A. Montagna, Ph.D.

Kevin Nelson, Ph.D. Candidate

Harte Research Institute for Gulf of Mexico Studies

Texas A&M University-Corpus Christi

6300 Ocean Drive, Unit 5869

Corpus Christi, TX  78412-5869

361-825-2040 Telephone

361-825-2049 Facsimile

Submitted To:

Coastal Bend Bays & Estuaries Program

1305 N. Shoreline Blvd., Suite 205

Corpus Christi, TX  78401

CBBEP Project Number: 2008-18

January 8, 2009

# Observation Data Model (ODM) For Rincon Bayou, Nueces Delta

## Introduction

Restoration of Rincon Bayou and marsh in the Nueces Delta began in 1994 with construction of a channel to divert fresh water into the marsh (BOR, 2000). The diversion was filled in 2000 but reopened in 2001. In 2007, a pipeline began to deliver water directly into Rincon Bayou from the Calallen Pool. Extensive monitoring of the Rincon Bayou area has taken place, first funded by the U.S. Bureau of Reclamation, and more recently funded by the City of Corpus Christi. The purpose of the freshwater inflow diversions has been to restore marsh function and is part of an overall strategy to manage environmental flows from the Nueces River water system to increase firm yield of the water supply. The Nueces Delta is one of only three places in Texas where the permit and State orders require environmental flows, making this an important public issue.

After 14 years of study and management, the Nueces Delta experience presents the best template on how freshwater inflows can be studied and managed in other parts of the state, and more broadly, the nation. Recently, the Coastal Bend Bays & Estuaries Program has embarked on a strategy to acquire wetlands in the Nueces Delta to further enhance restoration and protection of the marsh.

The current need is to determine if, and how, the diversions have restored marsh structure and function, how best to operate the diversions, and what the environmental benefits have been. This will fulfill an existing need of decision-makers as to how best to operate the water system. However, the first step would be to integrate the disparate types of data collected. Data is currently housed at four different institutions: the Center for Coastal Studies, Texas A&M University-Corpus Christi (TAMUCC); Conrad Blucher Institute, TAMUCC ; Harte Research Institute for Gulf of Mexico Studies, TAMUCC; and The University of Texas Marine Science Institute. As well as being held in four different physical locations, the data is in many different formats. Thus, the lack of access and understanding about formatting make it very difficult (meaning very expensive) to bring data together in a synthetic analysis. However, relatively new fields of science have emerged that can be used to solve the problem.

Ecosystem informatics is the process of blending ecosystem studies, computer science, and mathematics to further contributions in these fields, but also to facilitate ecosystem and/or natural resource management. Various observing agencies, universities, and other organizations collect data from many sources and organize this data using a variety of means. Access to the data is often limited to one desktop. When shared on multiple desktops, complicated schemes are necessary to ensure concurrency of the data. Limited access and varied platforms inhibit the ability to synthesize data, which is necessary to perform analyses and employ visualization techniques.

Spurred by low-cost, highly-available sensors, groups like researchers in the weather forecasting community (Plale et al., 2006), and the Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI, http://his.cuahsi.org/) (Maidment, 2005) have developed the cyber-infrastructure to coordinate data from various sources and provide data-driven tools for forecasting, modeling, and adaptive sampling. CUAHSI has developed a database schema, or observations data model (ODM), for the purpose of providing a standardized repository for historical and continuous observations of hydrological data as well as the accompanying metadata (Tarboton et al., 2008). Developed in concert with the ODM, WaterOneFlow, is a set of XML-based web services that provide common Internet-based access methods for online data sources (Valentine and Whitenack, 2008).

The CUAHSI model is designed for hydrologic observations. Is the model generic, and robust enough to encapsulate observations from outside the hydrological paradigm, such as the biological data from Rincon Bayou? Can ODM provide the necessary data and metadata structures to store and maintain biological data and laboratory-derived data? This goal of the present work is to explore these questions by extracting Rincon Bayou data from historical sources, transforming the environmental data into the ODM schema and instantiating web services for online access to the data.

## Materials and Methods

### Information System Requirements

The ODM requires a relational data base management system, SQL Server 2005 in this case. To run the CUAHSI web services, MS Internet Information Server and MS Visual Studio 2005 were necessary. All components were installed on a server running MS Windows Server 2003. A blank database ODM schema was downloaded from the CUAHSI website: http://his.cuahsi.org/odmdatabases.html. The blank schema was attached to a workstation running SQL Express using the SQL Server Management Studio. It should be noted that while the Express version of SQL Server was used to create the final synthesis product, tools found only in the Enterprise version of the SQL Server (i.e., SQL Management Studio and SQL Server Integration Services (SSIS)) were used in the process.

### Data Sources

Data sources for this project focused on Rincon Bayou in the Nueces River Delta and Nueces Estuary system in Corpus Christi, Texas. Researchers from the Center for Coastal Studies at Texas A&M University - Corpus Christi (CCS), Harte Research Institute for Gulf of Mexico Studies at Texas A&M University - Corpus Christi (HRI), and the University of Texas Marine Science Institute (UT) have collected environmental observations in the region since 1994. Each observation by every institution was made at a specific location, however latitude and longitude were not provided for all of the points in data collected along transects. Coordinates for these sites were obtained by interpolation. A variety of data management methods were employed by the participating groups, including MS Access databases, MS Excel spreadsheets, and SAS-based data files.

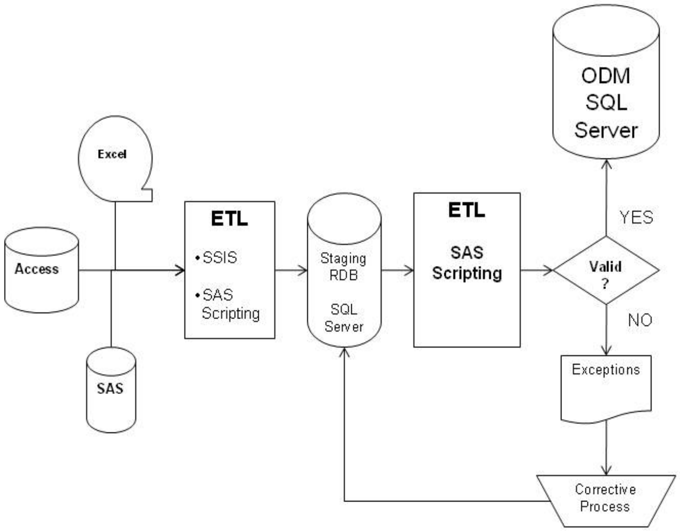

Fig. 1 - Extraction, transformation, and loading process

The overall process of data extraction, transformation, and loading (ETL) is illustrated in Figure 1. Extraction for the SAS-based and Excel-based data files was accomplished mainly through the use of SAS import functions (PROC IMPORT) and scripting using the SAS data language (SAS Institute, Inc., 1999). The SAS-SQL Server connection was accomplished using open database connectivity (ODBC) facilities in SAS, SQL Server, and Windows XP. Data extracted from original sources were staged on the SQL server, with little modification to their original structures or data types. Access database files were imported into the staging database using SSIS. Once staged, the SQL data language, via SAS PROC SQL, was used to transform, validate, and load the data into the ODM schema. Validation exceptions were routed to the data managers for research and correction and subsequent reloading. Once the ETL process was complete, the ODM instance was detached from the development workstation and reattached to the production server.

### Web Services

After the ODM instance was loaded, it was attached to the SQL server, and web services to access the ODM were created by tailoring the generic WaterOneFlow web services downloaded from the CUAHSI website at http://his.cuahsi.org/wofws.html. Installation proceeded as instructed by Valentine and Whitenack (2008). The procedure called for creating the required accounts on the operating system and SQL Server, mapping these accounts to logins and roles on the server, installing the web services application on the server and configuring the web server for access to the application.

## Results

Transformation of the data resulted in over 600,000 distinct observations in 44 different variable types grouped below into five general categories (Table 1). The number in parentheses denotes the number of distinct variables of that type. For instance, under biota, the variable “%Plant Coverage” denotes the percentage of cover for a specific plant species. In this exercise, there were 18 different species, including wrack represented in the data. Each species was coded as a separate variable. Similar schemes were used in the other biota variables accompanied by a number in parenthesis. For instrumentation, the number represents variables used in procedures used to derive a final observation. For instance, tare weight and rubble mass were used to derive sediment grain size percentages.

Table 1. Listing of variables in the ODM by general category. Numbers in parentheses represent the number of variables of a specific type.

| **Atmosphere** | **Biota** | **Water Nutrients** | **Water** | **Sediment** | **Other** |
|---|---|---|---|---|---|
| Barometric Pressure | Avian Activity | $[NH_4]$ | Dissolved Oxygen | $\delta\ ^{13}C$ | Days since precipitation |
| Cloud Cover | %Plant Coverage (18) | [N + N] | DO % Saturation | $\delta\ ^{15}N$ | Instrumentation (15) |
| Days since precipitation | Species (g/m$^2$) in core (7) | $[NO_3]$ | Color | %C | Tidal Stage |
| Precipitation | Species (mass) in core (7) | $[NO_2]$ | Redox Potential | %N | |
| Relative Humidity | Species (#/m$^2$) in core (7) | [Orthophosphate] | pH | %Sand | |
| Temperature | Species # in core (7) | | Depth | %Rubble | |
| Wind Direction | Species # indexed (5) | | SECCHI Depth | %Clay | |
| Wind Speed | [Chlorophyll] | | Salinity | %Silt | |
| Weather Conditions | | | Surface Appearance | | |
| | | | Specific Conductance | | |
| | | | Temperature | | |
| | | | Visual Turbidity | | |

The geographical extent of the synthesized data includes observation sites along the entire estuary (Figure 2). A total of 483 different sites are represented in the data. This includes interpolated positions along the plant coverage transects. Observations from Nueces Bay, Nueces River, and hypoxia studies in Corpus Christi Bay and Oso Bay are also included in the synthesis.

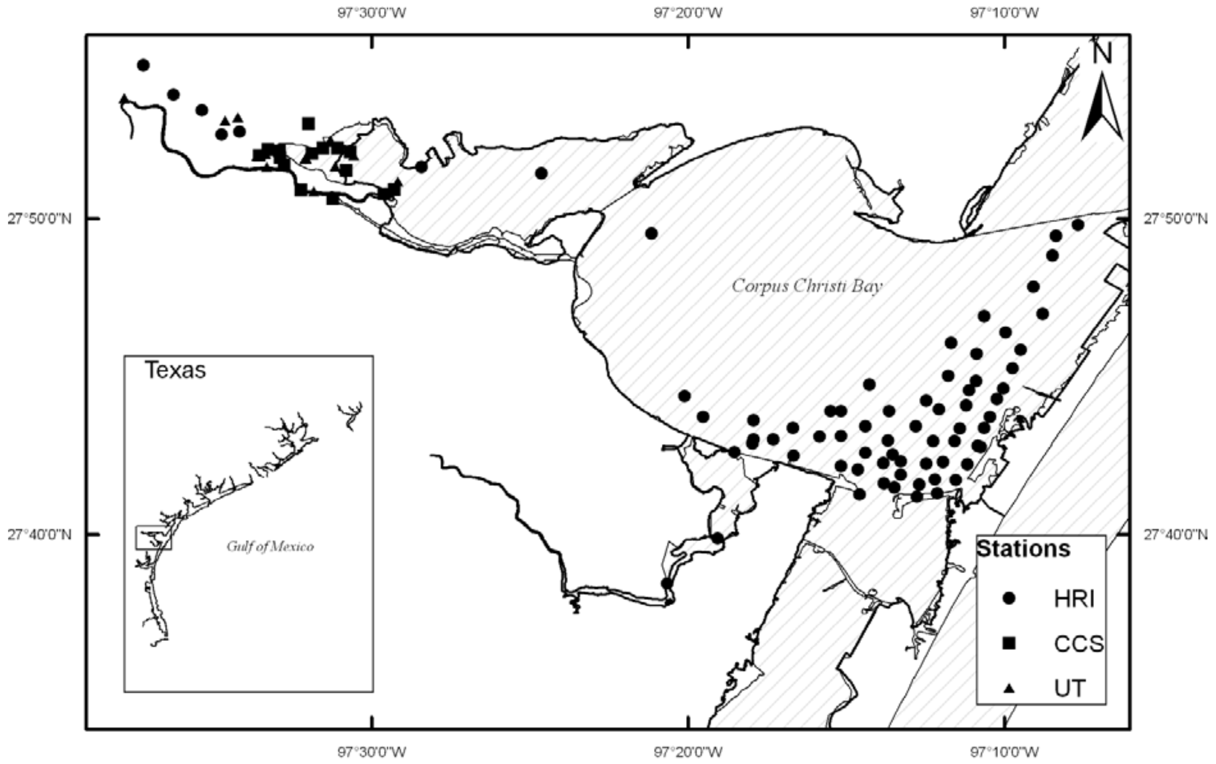

Figure 2. Geographical extent of synthesized data. Marsh plant transects shown as one point.

A description of the web services providing access to all synthesis data can be found at: http://ccbay.tamucc.edu/synth/cuahsi_1_0.asmx . The data can also be found, downloaded, and displayed from a website, http://ccbay.tamucc.edu/dash/ (Figure 3).

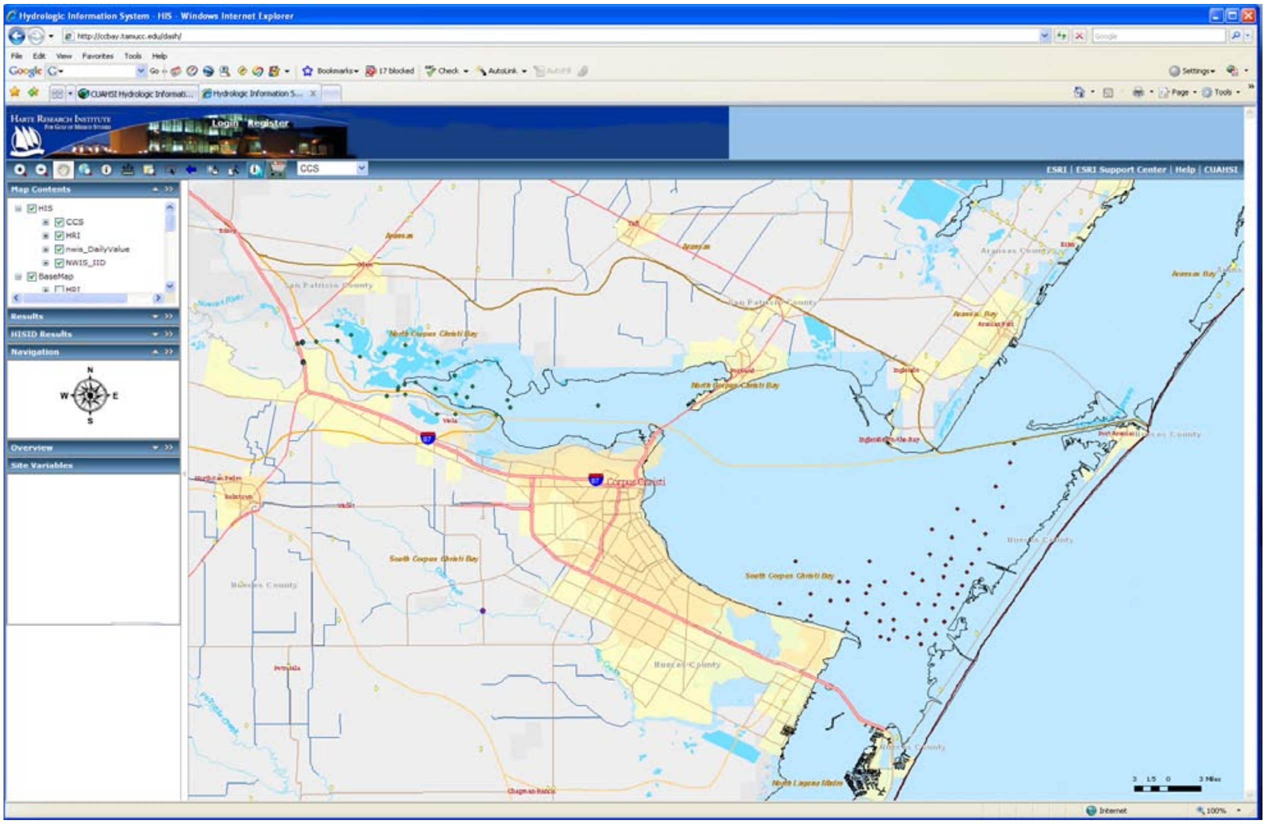

Figure 3. Image of the website (http://ccbay.tamucc.edu/dash/) that can be used to download and display data.

## Discussion

To examine the robustness of the ODM, a self-imposed restriction on ad hoc augmentation of the schema was used, meaning the ODM was not modified. However, some trivial modifications to the schema were made to facilitate loading of the data. For example, removing the restriction on the index field insertions in order to use externally created key field values. Some of the variable and method names from the Rincon Bayou data sources were not included in the original ODM controlled vocabularies. In these instances, variable and method names were inserted into the controlled vocabularies. Overall, these changes from the standard ODM structure are considered trivial because no new entities (i.e., tables) or relationships were defined.

The ODM was originally designed for hydrological data, such as simple time series observations from a single stream gauge. Series of this type in the synthesis, such as water temperature, were readily ported to the ODM and the metadata facilities were ample. However, more complicated samples were not straightforward. For instance, nutrient concentration data values result from multiple analyses performed on sample splits. Each analysis results in one data value, but all from the same sample. Logically, there is a one-to-many relationship between sample and data value. Constraints in ODM do not allow this relationship. In this synthesis, laboratory method metadata was loaded into the Methods table; a table more properly used for sample collection methods.

Each data value in ODM represents a measurement at some point in time and space. Date, time, latitude and longitude are each required fields. If the measurement is taken above or below the surface, then an additional offset value is required. Depth below the surface would be an example. ODM provides the flexibility to define this offset for other uses. For example, in this synthesis, the offset was used as key into a master species table. The variable “Species # indexed” is an example of this use of the offset. The count of a particular species is contained in the data value and a code for the species is contained in the offset. The master species table was not included in the synthesis as it would require the augmentation of the schema, therefore prior knowledge of the code is required. The alternative to this method would be to create a variable for each species. This solution could result in a proliferation of variables depending on the nature of the data. Where there is a small number of species represented in the data, a variable for each species might be more practical and in fact was used for the vegetation coverage in this synthesis. In some cases, both species and depth are needed to describe an observation. This situation occurs in species-based measurements in sediment cores. The offset denotes the section of the core. In this scenario, there is no place for the species code resulting in the need for a variable for each species. Fortunately in this synthesis, species were few, or represented at a taxonomic level high enough to result in few variables.

The nature of the transformation process removes some relationships implicit in the record-level collection of data. Transformation into ODM requires transposition and decomposition of the original record, (Figure 4). Typically, when hydrographic measurements are taken, values for water temperature, salinity, pH, and others are taken at the same time and depth. Logically, and practically, they are represented in a single record for each site, date, and depth. Indeed, all of the native data sources used in this synthesis employed this structure. Therefore, two records, each containing values taken at the same site, date, and depth can be considered replicates. Values in one replicate can be distinguished from those in another by their placement in the file. This is not a problem in most cases. After transformation into the ODM, the original record can be reconstructed by grouping the values by date, site, and depth. However, when there are replicate measurements this reconstruction is not possible without something denoting which replicate a data value belongs to. Facilities for easily denoting replicate associations are lacking. To differentiate replicates in ODM requires creating groups representing replicate numbers, REP1 for example, and keying each data value to the proper group.

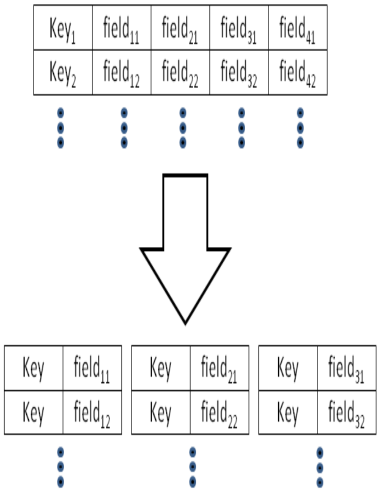
Key$_1$ field$_{11}$ field$_{21}$ field$_{31}$ field$_{41}$
Key$_2$ field$_{12}$ field$_{22}$ field$_{32}$ field$_{42}$
Key field$_{11}$
Key field$_{12}$
Key field$_{21}$
Key field$_{22}$
Key field$_{31}$
Key field$_{32}$

Figure 4. Decomposition of Records resulting from transformation into ODM.

In conclusion, the ETL was successfully used to load data into the ODM, and CUASHI web services were successfully used to provide public access to data. While the ODM is reasonably robust, there are some limitations in the ability to store laboratory method metadata and replicate sample numbers. One shortcoming of the ODM is the inability to store the hierarchy of biological names.

## References

Bureau of Reclamation (BOR). 2000. Concluding Report: Rincon Bayou Demonstration Project. Volume II: Findings. United States Department of the Interior, Bureau of Reclamation, Oklahoma-Texas Area Office, Austin, Texas.

Maidment, D., (Editor). 2005. Consortium of Universities for the Advancement of Hydrologic Science, Inc Hydrologic Information System Status Report, CUAHSI http://www.ce.utexas.edu/prof/maidment/CUAHSI/HISStatusSept15.pdf

Plale, B., Gannon, D. 2006. CASA and LEAD: Adaptive Cyberinfrastructure for Real-Time Multiscale Weather Forecasting. Computer 39:8, 56-64.

SAS Institute, Inc. 1999. SAS/STAT Users Guide, SAS Institute, Inc., Cary, N.C.

Tarboton, D.G., Horsburgh, J.S., and Maidment, D.R. 2008. CUAHSI Community Observations Data Model (ODM) Version 1.1 Design Specifications. CUAHSI http://his.cuahsi.org/documents/ODM1.1DesignSpecifications.pdf , 58 pp.

Valentine, D., Whitenack, T. 2008. HIS Document 4: Configuring web services for an observations database (version 1.0), San Diego Supercomputer Center, University of California, San Diego. http://his.cuahsi.org/documents/HISDoc4_ConfigureWebServices.pdf , 28 pp.