# TABLE OF CONTENTS

Table of contents ........................................................................................................... i
List of Figures ................................................................................................................... iii
List of tables ....................................................................................................................... iv
Acknowledgements ........................................................................................................... v
Abstract ............................................................................................................................. vi
1. Objective ....................................................................................................................... 1
2. Background and Rationale ........................................................................................... 1
   2.1: Rincon Bayou ......................................................................................................... 2
   2.2: Biological organization ......................................................................................... 3
   2.3: Ecosystem .............................................................................................................. 4
      2.3.1: Marine ecosystem ......................................................................................... 5
      2.3.2: Agro ecosystem ............................................................................................. 5
      2.3.3: Desert .............................................................................................................. 5
      2.3.4: Urban ecosystem ............................................................................................ 5
   2.4: Biological classification ......................................................................................... 6
   2.5: CUAHSI ............................................................................................................... 7
   2.6: DATA LOADER .................................................................................................... 10
3. Related Work ............................................................................................................... 12
   3.1: Texas water data from NWIS .............................................................................. 13
   3.2: National Ocean Data Center ............................................................................... 15
   3.3: Data Sources ....................................................................................................... 17
4. Expected Results ......................................................................................................... 18
5. Ecological Metadata Language (EML) ................................................................. 18

6. Design & Implementation .................................................................................. 22
   6.1: Introduction .................................................................................................. 22
   6.2: Procedures .................................................................................................. 23
   6.3: Development .............................................................................................. 26
   6.4: Variables Utilized ....................................................................................... 32
   6.5: Web services .............................................................................................. 33
   6.6: HydroDesktop ............................................................................................ 35

7. Future Work ........................................................................................................ 42

8. Conclusion ........................................................................................................... 43

Bibliography and References .................................................................................. 44
LIST OF FIGURES

1. Components of an ecosystem.................................................................................. 4
2. Biological classification system............................................................................... 6
3. Key components of CUAHSI-HIS........................................................................... 7
4. Screen shot of the HydroDesktop interface .......................................................... 9
5. Schematic of the HydroServer system ................................................................ 10
6. NWIS Mapper screen shot ..................................................................................... 14
7. Screen shot of the Deepwater Horizon map ......................................................... 16
8. Extraction, transformation and loading process.................................................... 24
9. Sites included in the project .................................................................................... 28
10. Christi Bay system stations so far included in the project ..................................... 29
11. UT Marsh database site example ......................................................................... 30
12. Database design overview .................................................................................... 34
13. Example showing downloads of data ................................................................... 35
14. Screen shot of the downloaded hydrological data ............................................... 36
15. Development approach for HydroDesktop .......................................................... 37
16. Example of Linnaean taxonomic system of classification .................................... 38
17. Screen shot of the IT IS search page .................................................................... 40
LIST OF TABLES

1. Data sources used for the current project ......................................................... 17
2. Listing of observation counts by source description ........................................... 27
3. Listing of top 25 sites by number of observations ............................................. 31
4. First 25 samples arranged by the number of observations ............................... 32
5. Operations provided by the web services ......................................................... 34
Acknowledgements:

Partial support for this project was provided by the Coastal Bend Bays & Estuaries Program through contract number 1105 to Dr. Paul Montagna, Harte Research Institute. Mr. Jace Tunnell of the CBBP provided many helpful comments during the development of the project. In addition, Mr. Kevin Nelson, Doctoral Candidate provided much help and guidance in the completion of this project.
The main purpose and objectives of this thesis are to create a database of environmental data for the Corpus Christi Bay ecosystem by compiling existing ecological databases; and to build user friendly tools to analyze and use the data and solve the issue of aggregation of biological data into levels of biological organization. The need to do all this was to solve the problem of displaying the data in the levels of biological classification. The database will be used to identify status and trends of selected parameters that indicate change in environmental quality in the Corpus Christi Bay region.
OBJECTIVE:

To develop an archiving and delivery system of existing water and sediment quality data in the Coastal Bend Bays & Estuaries Program (CBBEP) area by placing existing status and trends data in a modern developed database. The second objective is to update the existing web services to make that ecosystems database more user friendly.

BACKGROUND:

Around 15 years ago, the initial status and trends data report was completed [Ward and Armstrong 1997a, 1997b]. It contained analyses of water and sediment quality data through 1994. The status and trends project included accumulating and organizing the database as one level of achievement; and then analyzing the data and generating reports as another level. The issue here is that the complete database will have to be reorganized each time if the data is not forever archived. Lately, by using the Consortium of Universities for the Advancement of Hydrological Sciences (CUAHSI) Observation Data Model (ODM) [1], the Rincon Bayou and Corpus Christi Bay hypoxia databases have been archived.

Actually, the overall project goal is accomplishing the two levels (1) creating a database and developing tools for extraction, and (2) data analysis for trends. When complete, this will then be a database maintenance and management system that is both interactive and internet based. The ultimate database analysis product will be used for the identification
status of sub-areas of the bay system with water quality parameters surpassing water quality standards. By analyzing the water and sediment quality issues in the Coastal Bend to particular parameters investigated, such as bacteria and hypoxia, it will be possible to identify issues that will require management actions.

However, before analysis can begin, the first level (creating an accessible database) must be completed. Thus the goal of this thesis is to complete the database. Several problems had to be solved. One was how to deal with diverse data sources. Another problem was digitizing data. Another problem was dealing with biological data. With water quality data, there is one simple variable name and one simple value. This is not the case with biological data, which can be hierarchical because of levels of organization.

**Rincon Bayou:**

Rincon Bayou is a marsh system at the head of the Nueces River Delta. The marsh has degraded because of water diversions. To divert the fresh water into the Rincon Bayou and marsh in Nueces Delta in 1994, a new project was initiated for its development and restoration. A marsh [2] is a wetland submerged by water that lies generally around the water bodies like rivers and streams.

The commencement of the marsh restoration project was begun by constructing a channel that could divert fresh water into the marsh. The original goal of the diversion was to develop and restore the marsh function and monitor the environmental flows from the Nueces water system to the marsh.
After 14 years of study and management, the Nueces Delta experience exposed some guidelines of freshwater inflows for the adaptive management that was put in place in other parts of the Texas and also the United States in general [Montagna et al., 2009].

**Biological organization:**

Levels of biological organization are also called the hierarchy of life. The hierarchy goes from atoms to biospheres in which the upper levels will be known as ecological organization. The hierarchy goes as following (Wikipedia)

- Atom- smallest basic unit
- Molecule- group of atoms
- Organelle- group of molecules
- Cell- group of organelles
- Tissue- group of cells
- Organ- group of tissues
- Organ system- group of organs
- Organism- group of all lower level components
- Population- group of organisms
- Biocoenosis- group of populations
- Ecosystem- group of all biological environments and things associated with the environment
- Biosphere- group of ecosystems
Ecosystem:

An ecosystem [Fig. 1] is a biological environment that includes living organisms and non-living things or physical components such as sunlight, water, soil, etc., with which the organisms in a specific area interact [3]. An ecosystem could be temporary or permanent. The whole set of organisms living in an ecosystem is known as a community. In a classic ecosystem, plants and photosynthetic organisms (Wikipedia) are the source of food for and heterotrophic organisms, such as animals.

Fig. 1: Components of an ecosystem.
Some examples of ecosystems are:

**Marine ecosystem** is a part of the aquatic ecosystems. Lagoons, oceans, mangroves, coral reefs, deep sea, all come under the marine ecosystem. They are opposite to the freshwater ecosystem on the basis of salt constitution. Major part of the earth surface is covered by the marine ecosystem.

**Agro ecosystem** involves agricultural lands with living and non-living organisms that are associated with the area of the land. It is said to be a part of the conventional ecosystem and deals with everything related to agriculture considering human interference.

**Desert** is an area that does not very support plant growth due to extremely less precipitation levels of 400 millimeters. The areas with deserts are called water deficit areas. These can actually be classified into three types depending upon the levels of precipitation as true deserts with less than 250 millimeters, semi deserts with precipitation between 250 and 400 millimeters and deserts with 400 millimeters.

**Urban ecosystem** constitutes the world created by humans. It includes cities, towns and other urban places created by human beings. The remaining types of ecosystem are desert, forest, savanna, tundra, rainforest, etc.
**Biological classification:**

A biological classification [Fig. 2] is a mechanism to accumulate and classify organisms based on biological type [3]. There are nine levels in the hierarchy with species having the basic rank. Rank is the position in the hierarchy shown below.

![Biological classification system diagram](image)

**Fig. 2:** Biological classification system.

Example: Fish are members of the **Animalia** **Kingdom** (animals) and are also classified into the **Phylum Chordata** and Fish are further classified into the **Vertebrata** **Subphylum** [14].
The database schema used for this project is that developed by the Consortium of Universities for the Advancement of Hydrological Sciences (CUAHSI). Their Observations Data Model (ODM) provides a standard database structure for storing environmental observations and is an integral part for the CUAHSI [Fig. 3] Hydrologic Information System (HIS) [Tarboton, Horsburgh et al. 2008]. This model works well for hydrological data, such as water flow and elevation. However, the ecosystem database that has to be created and archived would contain all kinds of biological data. This biological data must be arranged into twelve levels in the ODM database according to levels in the ecosystem; so every living organism will have 12 names associated with each level.
The end user might want to analyze the data in every level or a specific level, which requires that the database be built by aggregating data accordingly in those twelve levels. So, tools must be developed so that when the user downloads required data, it gets downloaded in all the levels. For example, one must be able to retrieve the 12 names of an organism in 12 levels by searching by its common name. This could be achieved by actually giving identifiers to the names of each organisms and writing a code but that would not make it generic.

The CUAHSI organization provides users the options to find and get data, analyze data, model data, publish data, develop tools and develop tools. It includes components such as Web Services, HIS Central, HydroServer, and HydroDesktop.

To get, analyze, model the data, **HydroDesktop** is one of the tools provided by the CUAHSI-HIS organization.
Fig. 4: Screen shot of the HydroDesktop interface.

The HydroDesktop [Fig. 4] is a tool that is open source and can be used by the public. It is basically used to find, get and analyze the data as mentioned before. It goes hand in hand with the HydroCatalog for work in indexing the data to find where the data is and also to find whether it is available or not. After looking for the availability of the data using the HydroCatalog, it retrieves the data from the HydroServers. The data that is available in the HydroServers [Fig. 5] is the data published by the organization after much research and analysis and also the data provided by various organizations around the world.
The CUAHSI-HIS organizations supports the procedure to load data into its Observation data Model abbreviated as ODM. A tool called data loader is used to load data into the Observations data model.

**DATA LOADER:**

There is a tool that enables administrators of the local organizations to load data into the ODM [13]. This tool can be downloaded onto the system from the CUAHSI website. But the basic requirements of the system are it must have Windows XP, VISTA or Windows 7, and Microsoft SQL Server 2005 or higher must be installed before downloading and installing the data loader. The first version of the tool was data loader 1.1 and data loader 1.3 is the latest version.
The data loader has a sequence of steps, and a particular way in which the data must be loaded. It is not compulsory that every kind of data will be loaded by one particular organization. This order is important to avoid problems with the key (primary and secondary) constraints [13]. Each of the data must be loaded in a separate excel sheet and corresponding constrained naming conventions must be utilized. If a naming convention for a variable or a site name is unavailable, then a request can be placed to the organization to add it in the standard naming conventions list.

The following is the order in which data must be loaded:

1. Load Sites
2. Load Variables
3. Load OffsetTypes
4. Load ISOMetadata
5. Load Sources
6. Load Methods
7. Load LabMethods
8. Load Samples
9. Load Qualifiers
10. Load QualityControlLevels
11. Load DataValues
12. Load GroupDescriptions
13. Load Groups
14. Load DerivedFrom

15. Load Categories

For Example, when the “Sites” are considered, the fields that must be included in it’s excel sheet are SiteCode, SiteName, Latitude and Longitude.

RELATED WORK:

As a part of the related work, some ecological databases will be discussed in this section. This is necessary because the aspects that are lacking and important for a database of ecological systems will be understood. So, the databases that will be discussed are

1. Texas water data from NWIS: http://waterdata.usgs.gov/tx/nwis
3. Deepwater Horizon Spill
   Data: http://www.nodc.noaa.gov/General/DeepwaterHorizon/ships.html
6. Texas water data services: http://data.crwr.utexas.edu/wsdl.html
Texas Water Data from NWIS:

The NWIS corresponds to the National Water Information System. It belongs to the U.S Geological Survey (USGS). It is a diversified application that backbones for possessing, processing and archiving the water data. This application of the U.S Geological Survey is an open application and hence can be used by anyone connected to the internet.

This particular website provides various ways and options to analyze the water data based on the Data category and the Geographic area. The data category includes

1. real-time data
2. Site information data
3. Surface water data
4. Ground water data
5. Water quality data
6. Mapper data

The following the example for the mapper [Fig. 6]. A print screen of the website after selecting the mapper option is shown.
The above map shown depicts the different locations at which the experiments have been conducted. They are to be called sites in this context. These sites can be categorized into Surface-water sites, Groundwater sites, spring sites, Atmospheric sites and other sites. The markings of these sites are shown pictorially on a map.

If the real-time data option [4] is selected in the data category, the real-time water data is displayed on a map with colorful dots each depicting the percentile ranges. For example black indicates high, dark blue indicates greater than 90 percentile, lighter blue indicates 76th to 90th percentile and so on. This data is collected every 60 minute time interval and reported to USGS every one to four hours.

**Fig. 6:** NWIS Mapper screen shot.
National Ocean Data Center:

It is abbreviated to NODC [Fig. 7]. The National Oceanic and Atmospheric Administration are in charge of handling and controlling the national ocean data center.

The NODC provides options of using various kinds of data such as

1. Deepwater Horizon
2. Ocean Data Access
3. Ocean archive system
4. NODC Geoportal (beta)
5. Data formats and Codes
6. CD-ROMs and DVDs

The Deepwater Horizon further subdivides into various sections offering us data from sections such as Archived deepwater horizon data, ocean profile data, satellite data, fisheries information, oceans current data, climatology products, Coastal ecosystem maps, resource on oil spills, response and restoration, etc.
This flexible interface provides the option of downloading the map of the Gulf of Mexico that corresponds to the emergency fishery closures related to the oil spills. Any data available in these online sources need a format of the data to download that data. For example, this particular website requires some below mentioned files to access data from its websites. Those are .dbf corresponding to the attribute information, .lyr corresponding to the display of information, .prj corresponding to the projection information, .sbn corresponding to the spatial indexing, .sbx also corresponds to the spatial indexing, .shp corresponding to the geometrical information; it could be the information about the location, .shx corresponding to the indexing information.
DATA SOURCES:

Table 1: Data sources used for the current project

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Temperature and Salinity</th>
<th>Dissolved Oxygen</th>
<th>Suspended Solids</th>
<th>BOD and Related Parameters</th>
<th>Nutrients and Productivity Indicators</th>
<th>Trace Elements and Organic Pollutants</th>
<th>Bacteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FEDERAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US Geological Survey</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US Army Corps of Engineers</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US Env. Protection Agency</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOAA??</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USFWS??</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STATE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas Commission on Env. Quality</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Texas Dep. State Health Services</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Texas Parks and Wildlife Department</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Texas Water Development Board</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Texas AgriLife Extension Service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>LOCAL GOVERNMENT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nueces River Authority</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>City of Corpus Christi</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Nueces County</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>ACADEMIC/ NON-PROFIT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coastal Bend Bays &amp; Estuaries Program</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Texas A&amp;M University (CBI, TAMUCC, TAMU, CCS, HRI)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>University of Texas (UT, UTMSI, MA-NERR)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Whitledge (HRI-Montagna)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Table 1 shows the different sources of data from different organizations like federal, state, local government and non-profit organizations that are included as input to the database. It also shows the data types that are included in their data.

**EXPECTED RESULTS:**

A database is created by comparing and observing other existing ecological databases; user friendly tools to analyze and use the data will be built and the issue of aggregation of biological data into levels would be addressed. The work to achieve this is still in progress as the research is being done. So far, the aspects of biology required for the research have been studied. And various ecological databases have been discussed.

**Ecological Metadata Language (EML)**

EML is a standard for metadata developed for the field of ecology. It is mostly dependent on the effort from the Ecological Society of America (Michener et al., 1997). The Ecological metadata language is implemented as an array of XML document types that could be used in a modular and extensible way to collect ecological data [24]. A single module of the EML is used to sketch the particular logical part from the total metadata that needs to be added to the complete ecological dataset.
The following are the features of the Ecological Metadata Language:

1. **Compatibility**: EML metadata standard is adapted from other metadata standards and therefore will be well compatible.

2. **Modularity**: Instead of making it single entity standards, the EML is designed as a set of different modules that enable easy and better development of the language in all the angels.

3. **Detailed Structure**: The language provides sufficient amount of data to understand the processing and analyzing of the data.

4. **Strong Typing**: EML is implemented using the Extended Markup Language (XML) which is obtained from the World Wide Web Consortium and is a strongly typed language.

**Examples:**

**Ex 1: A valid EML document [24]:**

```xml
<?xml version="1.0"?>
<eml:eml
  packageId="eml.1.1" system="knb"
  xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 eml.xsd">
  <dataset id="ds.1">
    <title>Sample Dataset Description</title>
    <creator id="23445" scope="document">
```
Ex 2: An invalid EML document

<?xml version="1.0"?>
<eml:eml packageId="eml.1.1" system="knb"
xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1 eml.xsd">
<dataset id="ds.1">
<title>Sample Dataset Description</title>
</dataset>
</eml:eml>
Example 2 is invalid because of the usage of non-existent reference. The reference 23445 exists but the reference 23446 does not exist. Therefore the EML in example 1 is valid because it does not duplicate identifiers, does not have a non-existent reference nor a conflicting id attribute and a <references> element.
DESIGN & IMPLEMENTATION

INTRODUCTION:
Several organizations and agencies have been collecting water and sediment data from the Corpus Christi bay systems since many years now. The data types that have been accumulated by these several teams have been exceptionally good. To study the several evaluations of the estuarine processes, biological measurements in sediment and water column, the environmental variables and water characteristics and dynamics have all been accumulated.

The final goal of the current work is to save and archive this data into a database that is more users friendly which will enable the analyses of the current status and trends and recognize methods in the bay system for variability.

Another idea that could be added to this project or work would be to publish all the collected data and make it available to the scientists and also the general public who could utilize this data for the advancement of science. The Harte Research Institute’s assets have been utilized to deploy a high industry-standard database management system.

The data from the database is systematically backed up and is safely stored within the Texas A&M University-Corpus Christi firewalls. The database is kept restricted from direct public access and by using web services; it is alternatively made accessible through the Harte Research Institute’s web server.
Some software applications that enable us to discover and harvest the required data for the desktop computers have been recognized. These tools will help in querying the database, downloading the required data and then graphically representing the required data upon request.

**PROCEDURES:**

The data that is to be loaded into the database reside both in electronic and paper records. These records were not of acceptable quality since they were made lots of years ago and hence optical character recognition could not be utilized to electronize this data.

Therefore the data from the records was manually entered into the excel sheets so that it can be easily converted and imported into the database.

For those records that were electronic, the source database’s copy was obtained. Textual extracts from important databases like the Texas Parks and Wildlife Department (TPWD), Access databases, SAS tables, Excel sheets were used as the various types of source files for the expression of the data. Rare cases occurred when the definitions for these data sources were alike. So, it can be incurred that each specific sources of data were respectively considered leading to no or almost any code reusability.

For this project, the CUAHSI organization that expands to the Consortium of Universities for the Advancement of Hydrologic Sciences developed the database schema [Fig. 8] that has been mentioned before. For archiving the environmental observations, a standard database structure has been given by the ODM that expands to the Observations Data
Model which is the heart of the CUAHSI HIS i.e. Hydrologic Information System (Tarboton, Horsburgh et al. 2008).

**Fig. 8:** Extraction, transformation and loading process (Montagna & Nelson 2009).

The complete methodology of the extraction, transformation and loading is clearly shown in [Fig.8]. For specific format files, like the Excel-based and the SAS-based file, the extraction process was obtained primarily by the utilization of the SAS scripting and SAS import functions like Proc Import from the SAS data language.
SAS is a language that is used to process and execute data and helps in performing statistical analysis. SAS can be divided into two steps namely data steps and procedure steps. The data step deals with the initialization, handling and deleting the data while the procedure step deals with performing analysis and producing a desirable outcome.

**PROC IMPORT**

Example: PROC IMPORT OUT= WORK.auto1 DATAFILE= "C:\auto.xls"
   DBMS=xls REPLACE;
   SHEET="auto1";
   GETNAMES=YES;
   RUN;

The above is an example showing the format for Proc import. Proc Import is a function used to read/ write the excel files in SAS. In the above example, reading into the SAS is shown. The out option provides the flexibility to specify a name for that new file in SAS. Datafile option allows the user to provide the path for the particular Excel file. The DBMS specifies the kind of file being imported. Replace option will be used to overwrite the already existing file.

The connection establishment between the SAS and the SQL Server was achieved using ODBC that is the Open database connectivity option. So, data from the different sources were all mounted on the SQL server after alterations though little in the data types and structures of the original data.

Windows XP was used as a platform. The Access database files were incorporated into the database too. After mounting the data, SQL data language using SAS Proc SQL was
utilized for transformation, loading and validation of the data into the designed schema. After the completion of the extraction, transformation and loading process, ODM instance was separated from the workstation used for the development and linked to the production server.

The Microsoft SQL Server (MSSQL) database management system was used for the instantiation of the database schema developed by the CUAHSI. Because SAS is created to utilize MSSQL like a depository, using SAS scripting language (SAS Institute 1999) and SAS functions like PROC IMPORT, the input data files were incorporated into the SAS standard local structure. It is then converted into the Observations Data Model schema by utilizing the Microsoft Transact-SQL language. Few transformations were achieved with MS SQL Server Integration Services.

As already mentioned before, public is restricted from directly accessing the database server and is instead allowed to access the database through the CUAHSI web services. Using the user manual, the CUAHSI web services can be downloaded from www.cuahsi.org and installed (Valentine & Whitenack 2008).

**DEVELOPMENT:**

**Sites:**
The total number of observations so far by the transformation of the data which lie within the range of 1968 to 2011 counted to 8,938,576 from which around 7,504,629 [table 2] were from the TCEQ database. The total number of sites [Fig. 9] in the database is 8279 that are distributed all over the Texas state, bulk of these sites belonging to the TCEQ
database itself. Although, some sites [Fig. 10] did not belong to the Corpus Christi Bay systems, they were included in the transformation if these sites had observations of the environmental variables that are required in our database.

**Table 2:** Listing of observation counts by source description

<table>
<thead>
<tr>
<th>Source Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction from TCEQ database</td>
<td>7,504,629</td>
</tr>
<tr>
<td>Collection of Continuous sonde observations supporting multiple projects</td>
<td>1,060,998</td>
</tr>
<tr>
<td>Extraction from UTMSI Marsh Database</td>
<td>349,318</td>
</tr>
<tr>
<td>Collection of nutrient/chlorophyll concentrations for Texas Coastal waters</td>
<td>20,733</td>
</tr>
<tr>
<td>Hardcopy text of final report for Texas Water Development Board</td>
<td>2,898</td>
</tr>
</tbody>
</table>
Fig. 9: Sites included in the project.
Fig. 10: Christi Bay system stations so far included in the project.

In this project, the sites from the University of Texas Marsh database [Fig. 11] (Dr. Kenneth Du ton, personal communication) are exclusively given consideration. These sites mentioned are arrays of transects comprising of the several number of sample sites. The location of the transect array is represented by the latitude and longitude values available. Note that the latitude and longitudes give the location for the transect arrays but not the location of the samples. Each point on transects specified are determined by using the method of interpolation. So, the number of sites mentioned in Table-1 looks bigger than anticipated.
According to the count of the observations taken, the sites are arranged in the Table 2 as shown below. The table is updated from time to time by consistent observations accumulated by Dr. Paul Montagna’s work on the hypoxia in the Corpus Christi Bay area.

**Fig. 11:** UT Marsh database site example.
When these examples are studied, it can be inferred that one to five environmental observations are considered at a time. That is because each individual environmental variable is considered individually so that they could be counted even when they can be generalized under a bigger multivariable record.

**Table 3:** Listing of top 25 Sites by number of observations.

<table>
<thead>
<tr>
<th>Site Name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 24 in Bay/Project HY</td>
<td>182,986</td>
</tr>
<tr>
<td>Site 19 in Bay/Project HY</td>
<td>100,019</td>
</tr>
<tr>
<td>Site 2 in Bay/Project HY</td>
<td>78,931</td>
</tr>
<tr>
<td>Site 309 in Bay/Project HY</td>
<td>54,995</td>
</tr>
<tr>
<td>Site 41 in Bay/Project HY</td>
<td>53,684</td>
</tr>
<tr>
<td>Site 199 in Bay/Project HY</td>
<td>53,679</td>
</tr>
<tr>
<td>LAKE TRAVIS NEAR DAM AT LCRA TRAVIS COUNTY PARK</td>
<td>52,881</td>
</tr>
<tr>
<td>Site 39 in Bay/Project HY</td>
<td>52,643</td>
</tr>
<tr>
<td>Site 310 in Bay/Project HY</td>
<td>51,982</td>
</tr>
<tr>
<td>Site 430 in Bay/Project HY</td>
<td>47,362</td>
</tr>
<tr>
<td>Site 17781 in Bay/Project HY</td>
<td>40,222</td>
</tr>
<tr>
<td>Site 410 in Bay/Project HY</td>
<td>37,592</td>
</tr>
<tr>
<td>Site 17787 in Bay/Project HY</td>
<td>36,706</td>
</tr>
<tr>
<td>LAKE TRAVIS AT ARKANSAS BEND TO THE WEST OF RANCH ROAD 620</td>
<td>36,348</td>
</tr>
<tr>
<td>LAKE BUCHANAN NEAR BUCHANAN DAM APPROX 475 METERS TO THE WEST OF CORONADO RD</td>
<td>34,153</td>
</tr>
<tr>
<td>Site 17793 in Bay/Project HY</td>
<td>32,716</td>
</tr>
<tr>
<td>Site 440 in Bay/Project HY</td>
<td>31,511</td>
</tr>
<tr>
<td>Site 18247 in Bay/Project HY</td>
<td>30,944</td>
</tr>
<tr>
<td>Site 420 in Bay/Project HY</td>
<td>30,315</td>
</tr>
<tr>
<td>LAKE TRAVIS MID LAKE ADJACENT TO LAKEWAY/TO THE NORTH OF CORINTHIAN ROAD</td>
<td>29,257</td>
</tr>
<tr>
<td>LAKE LYNDON B JOHNSON NEAR ALVIN WIRTZ DAM APPROX 658 METERS NORTH OF FM 2147</td>
<td>26,728</td>
</tr>
<tr>
<td>RICHLAND-CHAMBERS RESERVOIR CHAMBERS CREEK ARM NEAR TCWCID</td>
<td>26,596</td>
</tr>
<tr>
<td>Site Name</td>
<td>Count</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1 PUMP STATION 570 M S AND 1.16 KM W OF INTERSECT OF SE 3240 AND SE 3250</td>
<td></td>
</tr>
<tr>
<td>LAKE BUCHANAN AT ROCKY POINT APPROX 1.3.KM NORTHWEST OF ROCKY RIDGE</td>
<td>26,254</td>
</tr>
<tr>
<td>LAKE TRAVIS MID LAKE AT CONFLUENCE WITH COW CREEK ARM AT PACE BEND APPROXIMATELY 2.02 KILOMETERS TO THE SOUTH OF FM 1431</td>
<td>23,574</td>
</tr>
<tr>
<td>INKS LAKE NEAR INKS DAM APPROX 161 METERS TO THE NORTHEAST OF ROY INKS DAM</td>
<td>22,794</td>
</tr>
</tbody>
</table>

**VARIABLES USED:**

There are around 119 variables that have been coded into the database so far. Among these 119 variables, lots of them are from the water quality and sediment data observations but the biota variables are also added along with them. Lot of the observations taken are from surface water but pore water and tissue, air, sediment are all included. Now Table 3 as shown below represents the first 25 variables arranged according to the number of observations of that particular variable.

**Table 4:** First 25 samples arranged by the number of observations

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>1,056,823</td>
</tr>
<tr>
<td>pH</td>
<td>852,000</td>
</tr>
<tr>
<td>Oxygen, Dissolved</td>
<td>835,054</td>
</tr>
<tr>
<td>Specific conductance</td>
<td>818,926</td>
</tr>
<tr>
<td>Oxygen, dissolved percent of saturation</td>
<td>286,112</td>
</tr>
<tr>
<td>Salinity</td>
<td>270,820</td>
</tr>
<tr>
<td>Phosphorus, total</td>
<td>243,781</td>
</tr>
<tr>
<td>Substance</td>
<td>Value</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Chloride</td>
<td>240,724</td>
</tr>
<tr>
<td>Sulfate</td>
<td>230,173</td>
</tr>
<tr>
<td>Chlorophyll a</td>
<td>228,330</td>
</tr>
<tr>
<td>Streamflow</td>
<td>224,602</td>
</tr>
<tr>
<td>Solids, total suspended</td>
<td>220,505</td>
</tr>
<tr>
<td>Nitrogen, NH3 + NH4</td>
<td>200,485</td>
</tr>
<tr>
<td>Coliform, fecal</td>
<td>195,633</td>
</tr>
<tr>
<td>Phosphorus, orthophosphate dissolved</td>
<td>195,466</td>
</tr>
<tr>
<td>Turbidity</td>
<td>164,002</td>
</tr>
<tr>
<td>Solids, total dissolved</td>
<td>155,595</td>
</tr>
<tr>
<td>Alkalinity, total</td>
<td>148,108</td>
</tr>
<tr>
<td>Carbon, total organic</td>
<td>144,101</td>
</tr>
<tr>
<td>Secchi depth</td>
<td>140,694</td>
</tr>
<tr>
<td>Nitrogen, nitrate (NO3)</td>
<td>132,405</td>
</tr>
<tr>
<td>Water Depth</td>
<td>129,683</td>
</tr>
<tr>
<td>Solids, volatile suspended</td>
<td>118,458</td>
</tr>
<tr>
<td>Nitrogen, nitrite (NO2) + nitrate (NO3)</td>
<td>116,642</td>
</tr>
<tr>
<td>Nitrogen, total kjeldahl</td>
<td>112,996</td>
</tr>
</tbody>
</table>

**WEB SERVICES:**

Because the data in the database is not readily available directly, it can be accessed through the web services. The web services are a dependable and convenient way to access the database [Fig. 12] by querying it and also displaying the data. Two procedures have been developed for the better access of the database namely HydroExcel and HydroDesktop.

Link, ccbay.tamucc.edu/CBBEPDAP_ODWSv11/cuahsi_1_1.asmx can be visited for new web services regarding this project. People can follow the link to access data from this database. Various operations offered by the web services used have been mention in the table shown below. More information about the operations can be obtained from the link specified above.
**Table 5:** Operations provided by the web services

<table>
<thead>
<tr>
<th>Operation</th>
<th>Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>GetSiteInfo</td>
<td>Site metadata</td>
</tr>
<tr>
<td>GetSiteInfoMultipleObjects</td>
<td>Metadata from multiple sites</td>
</tr>
<tr>
<td>GetSiteInfoObject</td>
<td>Site metadata (object form)</td>
</tr>
<tr>
<td>GetSites</td>
<td>Sites</td>
</tr>
<tr>
<td>GetSitesByBoxObject</td>
<td>Sites in a geographical box</td>
</tr>
<tr>
<td>GetSitesObject</td>
<td>Sites (object form)</td>
</tr>
<tr>
<td>GetValues</td>
<td>Values given site and variable</td>
</tr>
<tr>
<td>GetValuesForASiteObject</td>
<td>Values given site and variable (object form)</td>
</tr>
<tr>
<td>GetValuesObject</td>
<td>Values (object form)</td>
</tr>
<tr>
<td>GetVariableInfo</td>
<td>Variable information based on variable code</td>
</tr>
<tr>
<td>GetVariableInfoObject</td>
<td>Variable (object form)</td>
</tr>
<tr>
<td>GetVariables</td>
<td>Variable</td>
</tr>
<tr>
<td>GetVariablesObject</td>
<td>Variable (object form)</td>
</tr>
</tbody>
</table>

**Fig. 12:** Database design overview.
**HydroDesktop:**

The HydroDesktop is used in integration with the CUAHSI HIS and is a development made to help in recognizing, downloading and presenting the data. To utilize the application of HydroDesktop, users can connect to HIS Central. After the users get connected, they can freely access the data in several ways such as download the data, query the location and select data using sites and variables etc. An example for downloading data by selecting sites is shown below.

![Example showing downloads of data by selecting various sites](image)

**Fig. 13:** Example showing downloads of data by selecting various sites.

After accessing the data according to the way required, the data can be displayed graphically. An example [Fig. 13] screen shot from the UT Marsh has been shown. The HydroDesktop enables the users to plot the series of time from different origins of data in
a single graph. From the UT Marsh database, an example is shown in the Figure below obtained through the HIS central.

![Graph Image](image.png)

**Fig. 14**: Screen shot of downloaded hydrological data.

The HydroDesktop [Fig. 14] application can easily be downloaded from the CUAHSI website using the link,

[http://hydrodesktop.codeplex.com/wikipage?title=Getting%20HydroDesktop&referringTitle=Documentation](http://hydrodesktop.codeplex.com/wikipage?title=Getting%20HydroDesktop&referringTitle=Documentation). The manual is also readily available in the CUAHSI website specifying the methodology to use the application and access the data through the application.
HydroExcel is used to connect to the CUAHSI based web service. It is an application that specifically allows the user to connect to a web service in CUAHSI, access or download the data and display the data accordingly. It might seem similar to that of HydroDesktop but it is difficult to use and needs intense setup comparatively. A user can download the application at [http://his.cuahsi.org/hydroexcel.html](http://his.cuahsi.org/hydroexcel.html) along with the HydroObjects at [http://his.cuahsi.org/hydroobjects.html](http://his.cuahsi.org/hydroobjects.html) which supports the HydroExcel application. The manual for both of these is also available in the website.

![Diagram of HIS Server Distributed Water Data Services and National Dataset Water Data Services](image)

**Fig. 15:** Development approach for HydroDesktop ([http://hydradesktop.codeplex.com](http://hydradesktop.codeplex.com)).
Biological classification and data Organization:

The third objective of the thesis paper was to create a database that is organized in such a way that when a particular data is tried to be retrieved, then, output will be in the levels of biological classification. The levels are Species, Genus, Family, Order, Class, Phylum, Kingdom, Domain and Life.

<table>
<thead>
<tr>
<th>Common Name:</th>
<th>Human</th>
<th>Nine-Banded Armadillo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom:</td>
<td>Animalia</td>
<td>Animalia</td>
</tr>
<tr>
<td>Phylum:</td>
<td>Chordata</td>
<td>Chordata</td>
</tr>
<tr>
<td>Class:</td>
<td>Mammalia</td>
<td>Mammalia</td>
</tr>
<tr>
<td>Order:</td>
<td>Primata</td>
<td>Cingulata</td>
</tr>
<tr>
<td>Family:</td>
<td>Hominidae</td>
<td>Dasypodidae</td>
</tr>
<tr>
<td>Genus:</td>
<td>Homo</td>
<td>Dasypus</td>
</tr>
<tr>
<td>Species:</td>
<td>Homo sapiens</td>
<td>Dasypus novemcinctus</td>
</tr>
</tbody>
</table>

Fig. 16: Example of Linnaean taxonomic system of classification.

Coming to the biological classification, it is important because it is easier for biology analysts around the world to understand the difficulty with the diversities of biology. More than a million organisms have been discovered so far and there are millions of more organisms that are yet to be discovered. For this purpose, taxonomy or biological taxonomy was used. Taxonomy is the kind of process that is used to recognize, name and organize the discovered species. Every level in the classification can be referred to as a rank. Species level means a set of organisms grouped together based upon several characteristics that they possess.
So, every living organism that has been discovered by scientists from their discovery will be given a scientific name based on its category and characteristics. Therefore when a database is created, the data in the data has to be organized in such a way that when a data is requested or queried from the database, the names of that particular organism has to be given in various levels specified.

The problem is that there are some websites that return the scientific name for the name of the organism entered but how would the user know which level it belongs to. And also it is difficult to perform summary analysis on the data due to the problem specified above. Existing options to resolve the problem mentioned, though not complete but provide a sufficient path are ITIS, EOL, GBIF, Darwin core, etc. Darwin core provides a body of standards that help in sharing the biological data in generic form. Each of the organizations mentioned above are discussed as follows.
**Fig. 17:** Screen shot of the ITIS search page.

ITIS [Fig. 17] is abbreviated as Integrated Taxonomic Information System. It is an organization that provides a user interface to search for taxonomic information about living organisms. This is an interface to a database that provides dependable data on the names of species and other related hierarchical classification. The organization updates its database periodically to ensure the availability of the required information. But this is not satisfactory because when the name of an organism is given, it returns the available names of the species but in levels.

EOL is abbreviated as Encyclopedia of Life. It is an organization that provides a very interactive interface that helps the user acquire information regarding biodiversity. It provides options specifying the way for discovery of diversity, naming new species,
taxonomy information for that specific species, etc. The EOL still tries to organize data based upon the priority of the species.

The Catalogue of Life is another organization that works in partnership with the ITIS and possibly has its own interface that gives an option to search for the name of any organism. That means it allows the user to select the type of name that he requires like its common name, family name or in most of the cases, the species name. All the above mentioned organizations have been collecting data and archiving it into their designed frequently updated databases. User friendly interfaces to help access the databases are provided along with advanced search options. So, efforts have been made to make the biological data available to the scientists all over the world but not with a complete satisfaction of exact required data analysis.
CONCLUSION & FUTURE WORK:

Phase-2:
As discussed earlier, the actual project has two phases. The first phase of the project has been discussed and shown in the current thesis. The second phase of the project is to perform some analysis on the data and write a report that recognizes the regions with parameters; parameters that are trending lower, higher or exceeding standards. Both of these phases are actually considered as two different projects because of the amount of time it would require.

Though this is the second phase of the project, it is being done in parallel with the first phase by Mr. Terry Palmer under the supervision of Dr. Paul Montagna. Reporting is a part of this phase and the plan is to provide three types of reports namely a Quality Assurance Project Plan (QAPP), Quarterly reports and then the final report.

The goal of the project is to contribute in achieving the CBBEP Bays Plan Objectives WSQ-1 and WSQ-4. For the CBBEP, the aim of this phase is to analyze the data mined for the advancement of the phase-1. After the analysis is done, reports will be provided on the historical data from the CBBEP area.

Approaches to Statistical Data Analysis:

To analyze the data and also examine the hypothesis, several statistical tools will be used. Univariate and multivariate statistical tools will be used.
Database proposed in phase 1 of the project is created using the sources specified and the corresponding web services to access the database is in progress. These web services will help scientists and other users across the world to access the data and use it for further analysis. After both the phases, phase-1 and phase-2, the overall result of the project is to help CBBEP in its development to focus on water quality and sediment practices in regions affected by trends specified; And also to help the users (scientists working on water and sediment data, etc) to better address the issues concerning the water quality and sediment data regarding the bays, estuaries and marshes. And the problem of arranging data in the database according to biological classification and similar retrieval id discussed.
Bibliography & References:

1. http://his.cuahsi.org/
2. http://library.thinkquest.org/05aug/00690/PagesM/what_is_a_marsh.html
11. http://datadryad.org/


