DEVELOPMENT OF A DATA MANAGEMENT FRAMEWORK IN SUPPORT OF SOUTHEASTERN TIDAL CREEK RESEARCH

Danna Wolf , David L. White, Dwayne E. Porter, Denise M. Sanger, George H.M. Riekerk, Guy DiDonato, A. Fredrick Holland, David Dabney

AUTHORS Danna Wolf Data Management Lead 331 Fort Johnson Rd. Charleston, SC 29412

REFERENCE Proceedings of the 2008 South Carolina Water Resources Conference, held October 14-15, 2008, at the Charleston Area Event Center

Abstract. The NOAA Center of Excellence for Oceans and Human Health Initiative (OHHI) at the Hollings Marine Laboratory (HML) is developing a data management framework that supports an integrated research program across scientific disciplines. The primary focus of the database is to support environmental research focused on tidal creek watershed systems. Specifically, the current data holdings are derived from several state and federal research programs and integrated into a common database model to support current research being conducted under the OHHI program at HML. The Tidal Creek database was developed with the intent to support a well documented and open system. The result is a semantic database framework with descriptive ancillary data at the record level including methods, investigator names, date, locations and other descriptive elements. The primary users of the database are project personnel to meet analytical needs. The database is also available through a number of web-based applications that are designed to give users the necessary information to evaluate and access data. In addition, data can be accessed with Open Geospatial Consortium (OGC) standards, and species records and abundances are being made available to the Ocean Biogeographic Information System (OBIS). Overall, the Tidal Creek database summarizes the response of tidal creeks and watersheds to coastal development, and serves as a repository for environmental, demographic, and socio-economic data in the Southeast.

Introduction

The Hollings Marine Laboratory (HML) in Charleston, SC, USA is a recent addition to the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Coastal Ocean Science. The laboratory is a product of a long-term joint agreement among NOAA, the National Institute of Standards and Technology (NIST), the South Carolina Department of Natural Resources (SCDNR), the College of Charleston (CofC), and the Medical University of South Carolina (MUSC). This unique partnership allows basic, applied, and biomedical research expertise to provide science and biotechnology applications to sustain, protect, and restore coastal ecosystems, emphasizing linkages between environmental and human health.

The growing research focus on the interactions between ecological health and human health is creating exciting opportunities and challenges in the data management field. Thus, the focus is not limited to databases within a single scientific domain (e.g., estuarine ecology or animal physiology) but to develop data and information systems that will aggregate across scientific domains. This requires that the underlying database model is flexible and designed to meet current and potential future research directions, and can support multiple tasks ranging from data archive to dissemination in a spatial context.

Within the data management group at HML, we have developed the Tidal Creek database for southeastern US tidal creek data from 1994 to the present. These data span multiple state and federally funded initiatives that address the effects of land use on tidal creek ecology. In addition, NOAA's recent focus to better address humans as integral to ecosystems, and the human dimensions of ecosystems has focused these efforts to develop a database model that would support an integrated ecological research program (NCCOS 2007). Thus, the Tidal Creek database was developed with a focus on the use of Open Source and Open Standard technologies with the intent of providing data into larger federated systems. Due to the diversity of the data that would be stored and made accessible to several user groups, a second critical component of this database was to support thorough documentation of the analytical results at the record level. This paper documents the development of an integrated data management framework to support the following activities 1) the consolidation of several databases that supported previous southeastern tidal creek research programs; 2) support current and future data collections including data related to human dimensions research, and make data available to research staff for analytical processing; and 3) support data dissemination through the use of web technologies and open standards to non-HML OHHI users.

Background

Before development of the Tidal Creek database, data management activities were limited at HML with activities relegated to individual laboratories and researchers. The HML Tidal Creek database was initiated with the consolidation of several Microsoft Access™ databases from research projects conducted in South Carolina, USA dating back to 1994. The overall focus of these studies was to investigate the impact of land use change in tidal creek watersheds and the ecological changes within those systems. These projects were quite extensive in scope and included multiple principle investigators addressing several research hypotheses. This data management effort was initiated in support of the OHHI at HML a funded effort that was established within NOAA by Congress in 2003 that is largely the result of several National Research Council reports (NRC 1999; 2001; 2002). The inclusion of the previously identified research project data served two purposes: 1) to support research efforts within the OHHI by making available historical data that could be included in addressing project hypotheses; and 2) providing a foundation for the development of a database to support potential long-term data management activities for tidal creek data in the southeast.

A Stressor-Exposure-Response Model was developed under the framework of the US Environmental Protection Agency (EPA) Ecological Risk Assessment Model (EPA 1993; 1994). This model describes and explains the connection among human population density (the stressors), physical/chemical changes in the environment (the exposures), and the response among biological populations (Lerberg et al. 2000; Holland et al. 2004). Research focused on multiple environmental scales and examined land use patterns in creek watersheds, water quality, pathogens, chemical contaminants, and kinds and abundance of aquatic life in creek water and sediments. The primary parameters that make up the Tidal Creek database include physical water quality (salinity, temperature, dissolved oxygen, pH), nutrients (inorganic nitrogen and phosphorous), chlorophyll *a* biomass, an analytical suite of over 100 chemical analytes, species counts including benthic and nekton assemblages with their associated taxonomic structure, and bacteria and viral pathogens. In addition, watershed scale parameters include watershed boundaries, and within those defined boundaries are land cover, percent impervious cover, demographic (e.g., human population density), socio-economic and parcel data. The structure of the database is flexible and will support various levels of spatial and temporal queries from a single site to multiple sites within multiple watershed systems.

Due to the goals and integrated nature of the database it was determined that detailed metadata should be part of any returned record from the database. Efforts were made to develop the database structure to support specific metadata elements that transcend all research programs. Thus, project metadata was developed using the Dublin Core an ISO 15836:2003 metadata standard that provides a succinct description of the projects. In addition, metadata elements at the record level include contact information for primary researchers, analytical laboratories, data collectors, collection dates, geographic locations, analyte units and instrumentation detection units taking into account the FDGC Content Standard for Digital Geospatial Metadata (CSDGM) (FGDC-STD-001-1998) which is a federal mandatory metadata standard for federal geospatial data. It is important to recognize that although the same analyte types are found from multiple research programs different collection and analytical methods could negate the comparison of such analytes. Therefore, the goal of this database is to provide all of these metadata elements at the record level so that every analyte has a history that can be compared against other records in the database. Researchers must understand the history and context of data collections to accurately analyze combined data in a meta-analysis or risk the introduction of error. In addition, thorough documentation using existing metadata standards will support archival and increase the data lifecycle.

Database Design

The overall structure of the database is primarily realized with the use of three general table types including attribute, reference, and data. The overall structure of the database is based on the concept of the data collection that occurs in tidal creek environments. The collection in this case is viewed as a field data collection where multiple parameters are collected. Each collection is assigned a *collection ID* before the field expedition. These collections are then subdivided for analytical purposes and quality assurance (QA). Uniqueness at the result level is maintained by using constraints; every result must have a unique combination of fields including collection ID, parameter ID, lab replicate, method ID and unit. Collections have certain properties such as collection method, station, latitude and longitude (bounding box is optional), date and time, tidal stage, personnel, the name of the project that it is part of, and others. The *Collection* table is viewed as an attribute table because it provides attributes about the collection that detail the context of a result.

The *Parameter* table is another example of an attribute table. After initial design of the database and the completion of data upload and integration of the project databases, the result was a mismatch of parameter names and units. The parameter table was developed to consolidate, when appropriate, parameter names and units. The result is a higher level documentation of all the parameters currently in the database. This provides a succinct way to document all parameters and provide descriptions. We have also leveraged this table to allow other federal databases to serve as reference sources for the analytes understudy. For example, many of the organic chemical analytes are documented by the NIST with a CAS Registry Number and ancillary information in web accessible systems. The CAS Registry Numbers are stored in the *Parameter* table. This same concept is used for species names (common and scientific) that are maintained by the Integrated Taxonomic Information System (ITIS) where the *Species* table leverages the taxonomic serial number from the ITIS. The result is that applications can be developed that leverage other authoritative distributed databases to verify naming and taxonomic nomenclature that either link potential users to these knowledge systems and/or verify database records.

Reference tables are intended to limit character data types repeatedly entered as fields in the various tables. Two primary reference tables were implemented including *Code_Table* and *Table_Entry* to manage an unlimited number of references for basic descriptive data. Human data entry error can become a significant problem in an information rich descriptive database. Using a reference in this case *Table_Entry,* a creek name is entered once and defined. *Table_Entry* serves as a catch all for the various descriptive elements that result from data collections and analyses of environmental parameters that range from physical attributes of the landscape to taxonomic nomenclature. *Code_Table* is a way to logically group sets of *Table_Entry* records and can be used to filter and select desired values. The individual records that makeup *Table_Entry* can be easily grouped into broad categories such as creek names that are defined in *Code_Table* under a single category. The end result of this design is that overall database structure remains relatively small without the need to add additional tables for specific object groups such as a creek name table or tide stage, etc.

Data tables are based on the concept of a collection and the corresponding physical, chemical and biological results. The result data are entered into five primary tables that include *Exposure_Data, Response_Aggregate,* R*esponse_Individual, Water_Quality, and System_Attributes.* HML defines **exposure** data as the physical and chemical changes in the environment that affect biological processes in an ecosystem. HML defines **response** data as the biological changes in the ecosystem that are responding to changes in the physical and/or chemical environment. Records in the *Exposure_Data* table are unique and documented with a numerical or text result value, a parameter type, units, analytical method, detection limit (if applicable) and others. The *Response_Aggregate* and *Response_Individual* tables are specific to composite and individual animal records and record data such as chemical concentrations and lipid analyses from animal tissues. The *Water_Quality* table is a result of continuous data sonde deployments in study creeks. *System_Attributes* addresses the need to capture data at a system or watershed scale. Data are generated from several sources and include demographic, socio-economic, land use, impervious cover, and county parcel data for each system under study. These data are calculated within a desktop GIS and analytical outputs are entered into the database as tabular data.

Data access and retrieval

A critical component of the Tidal Creek database is to make access as seamless as possible in support of HML and OHHI researchers. A second goal was to develop a system that could support data dissemination using web technologies. HML Data Management staff worked with researchers to understand access needs and types of queries that would be needed to support data analyses. The primary requirement for the database model was to support the Stressor-Exposure Response Model (e.g., Holland et al. 2004) and support outputs to statistical analysis programs. However, different sets of requirements were needed to make data available through web applications and web services. The result supports multiple uses including use by researchers, dissemination via the web, and an archival system for ongoing data collections.

To meet the needs of the data users within HML, MS Access™ is used to connect with the database with Open Database Connectivity (ODBC) drivers. The benefits of this have been invaluable as it provided a standard interface for researchers to interact with the database in a familiar environment. Complicated SQL queries can be written by data management staff and sent as text over e-mail and immediately used by researchers within MS Access™. The built-in form development tools in MS Access™ are easily deployed and made available to all users from a single *intranet* location. In addition, since data are generally entered into MS Excel™ these data can be imported into MS Access™ and uploaded into the database.

The Tidal Creek database was developed to not only support research analyses and program needs at HML, but to make data available to user communities outside of HML. A suite of web applications was developed to provide a user the means to explore and download data from the Tidal Creek database. A second objective of the data dissemination activities was to make data available to users using web services and to push data into larger federated data systems. MapServer serves as an OGC compliant application that supports data dissemination with the Web Feature Service (WFS) and Web Map Service (WMS) standards. In addition, Tidal Creek database species records are being pushed to OBIS an online database supporting global marine species records.

Summary

The Tidal Creek database was designed to facilitate research, data archival, and dissemination in an interdisciplinary environment. A critical problem was to develop a framework to support data upload, management, and dissemination in a consistent manner across the interdisciplinary research staff at HML. In addition, with the inclusion of past research data and the intent to make data available with web technologies users need to have access to metadata to properly evaluate data records. This was realized with the consistent generation of metadata using common metadata standards as an integral component of the database model. As web technologies continue to advance and more data resources come on-line there will be an increased focus on integrating various disparate data sets to address environmental research issues related to human and ecosystem health. The Tidal Creek database can be accessed through the HML web site (http://hml.noaa.gov/).

Acknowledgements

This work contributes to the NOAA Center of Excellence in Oceans and Human Health at Hollings Marine Laboratory. This work was conducted with the Ecological Research, Assessment and Prediction Group at Hollings Marine Laboratory (HML). HML is a partnership laboratory operated by NOAA for NOAA, NIST, College of Charleston, Medical University of South Carolina, and South Carolina Department of Natural Resources scientists to work together towards an interdisciplinary understanding of coastal ecosystem health. NOAA is an agency of the US Department of Commerce. Mention of trade names or commercial products does not constitute endorsement or recommendation for their use by the United States government.

Literature Cited

Environmental Protection Agency. (1993). *A review of ecological assessment case studies from a risk assessment perspective*. (EPA/630/R-92/005). (US Environmental Protection Agency, Washington, DC)

Environmental Protection Agency. (1994). *Peer review workshop on ecological risk assessment issue papers*. (EPA/630/R-94/008). (US Environmental Protection Agency, Washington, DC)

Holland, A. F., Sanger, D. M., Gawle, C. P., Lerberg, S. B., Santiago, M. Sexto, Riekerk, G. H. M., Zimmerman, L. E., & Scott, G. I. (2004). Linkages between tidal creek ecosystems and the landscape and demographic attributes of their watersheds. *Journal of Experimental Marine Biology and Ecology*, 298, 151-178

Lerberg, S. B., Holland, A. F., & Sanger, D. M. (2000). Responses of tidal creek macrobenthic communities to the effects of watershed development. *Estuaries*, 23, 838-853

National Centers for Coastal Ocean Science. (2007). *National Centers for Coastal Ocean Science human dimensions strategic plan (FY2009-FY2014).* (National Oceanic and Atmospheric Administration, National Ocean Service, National Centers for Coastal Ocean Science, Silver Spring, MD)

National Research Council. (1999). *From monsoons to microbes: Understanding the ocean's role in human health.* (National Academy Press, Washington, DC)

National Research Council. (2001). *Climate, ecosystems, and infectious disease. committee on climate, ecosystems, infectious disease, and human health. board on atmospheric sciences and climate division on earth and life studies.* (National Academy Press, Washington, DC)

National Research Council. (2002). *Problems, promise and products.* (National Academy Press, Washington, DC)