A COASTAL WATER QUALITY METADATA DATABASE FOR THE SOUTHEAST U.S.A.

Wade M. Sheldon, Christine Laporte, Travis Douce, and Merryl Alber[4]

AUTHORS: Department of Marine Sciences, University of Georgia, Athens, Georgia, 30602.
REFERENCE: *Proceedings of the 2011 Georgia Water Resources Conference*, held April 11-13, 2011, at the University of Georgia.

Abstract. This paper describes the development and initial implementation of a Coastal Water Quality Monitoring Metadata Database for the Southeast region (from NC to FL), which was developed with funding from the National Park Service. The database was designed to store detailed information on water quality monitoring programs operated by federal, state and municipal agencies, as well as by research institutions, including monitoring station locations, measured parameters, program contacts, and links to program web pages and data downloads. Water quality parameter records are classified into parameter groups and categories to support searches at varying levels of specificity, and are matched to US EPA STORET codes when possible for interoperability with federal databases. Fields for defining sample media, units and methodology are also provided for additional context. A prototype web portal, web services and mapping services were developed to support search and display of database contents, and to support leveraging by other database and portal efforts. Information from 41 monitoring programs in the South Atlantic was initially loaded into the database in 2009, including metadata on 16,182 stations at which 1093 distinct parameters are measured; the number of programs is currently being expanded. This database provides an ongoing inventory of monitoring activities for the southeast region and will help to facilitate identification of data gaps or under- or over-sampled areas. On a broader scale, the project's water quality metadata database and web portal have timely relevance to the broad community of coastal managers, researchers, planners and constituents as they make significant progress in leveraging and focusing regional associations and partnerships.

INTRODUCTION

Numerous agencies and institutions collect water quality data in the southeast coastal region of the U.S. In 2007, the National Park Service Southeast Coast Network (NPS SECN) identified the need to consolidate this regional information so that coastal managers, researchers and other users can readily identify sampling efforts by location, parameters measured, and methodology. NPS requested the development of a water quality monitoring metadata database for the southeast that would serve as a gateway to this information, and the Georgia Coastal Research Council (GCRC), located in the Department of Marine Sciences at the University of Georgia, was funded to develop this resource (i.e. the SE Coastal Water Quality Monitoring Metadata Database). Representatives from various agencies and institutions active in coastal water quality monitoring provided guidance on the scope and functionality of the proposed database during a 2008 workshop convened at the Hollings Marine Laboratory in Charleston, SC. An initial metadata database was then developed and deployed in 2009 (Sheldon et al., 2009). This paper provides a brief overview of the structure and content of the database, as well as web-accessible user interfaces, future development plans, and applicability to resource managers and environmental scientists in Georgia and across the SE USA.

METHODS

Database Design

Prior to designing the database we conducted a search to identify suitable metadata content standards and data models to leverage. We found that most published database designs and metadata standards are oriented towards documenting measurement details, primary data collection activities and data set characteristics rather than monitoring programs and locations. However, portions of the Environmental Sampling, Analysis and Results Data Standards from the Environmental Data Standards Council were suitable, particularly the Project (EDSC, 2006a) and Monitoring Location (EDSC, 2006b) metadata standards.

Using the EDSC content standards as a guideline we designed a relational database, "NPS_WaterQuality", to store all metadata needed for the project using Microsoft SQL Server™ 2000. The database model was implemented using a fully-normalized table design with referential integrity constraints (i.e. foreign key relationships) to prevent entry of incomplete or unlinked information (Riordan, 1999). Select portions of the database model are illustrated as entity-relationship diagrams below (figs.1-2), and the complete design is described in detail in Sheldon et al., 2009.

Based on initial investigations, we found that available information about monitoring activities varies greatly in scope and detail. In order to provide the needed flexibility in the database model, we used a two-table design for both organization and monitoring program metadata. The primary tables (Organizations and Monitoring Programs) contain fixed fields with appropriate data types for storing information common to all organization and monitoring program entries (name, acronym, description, web site

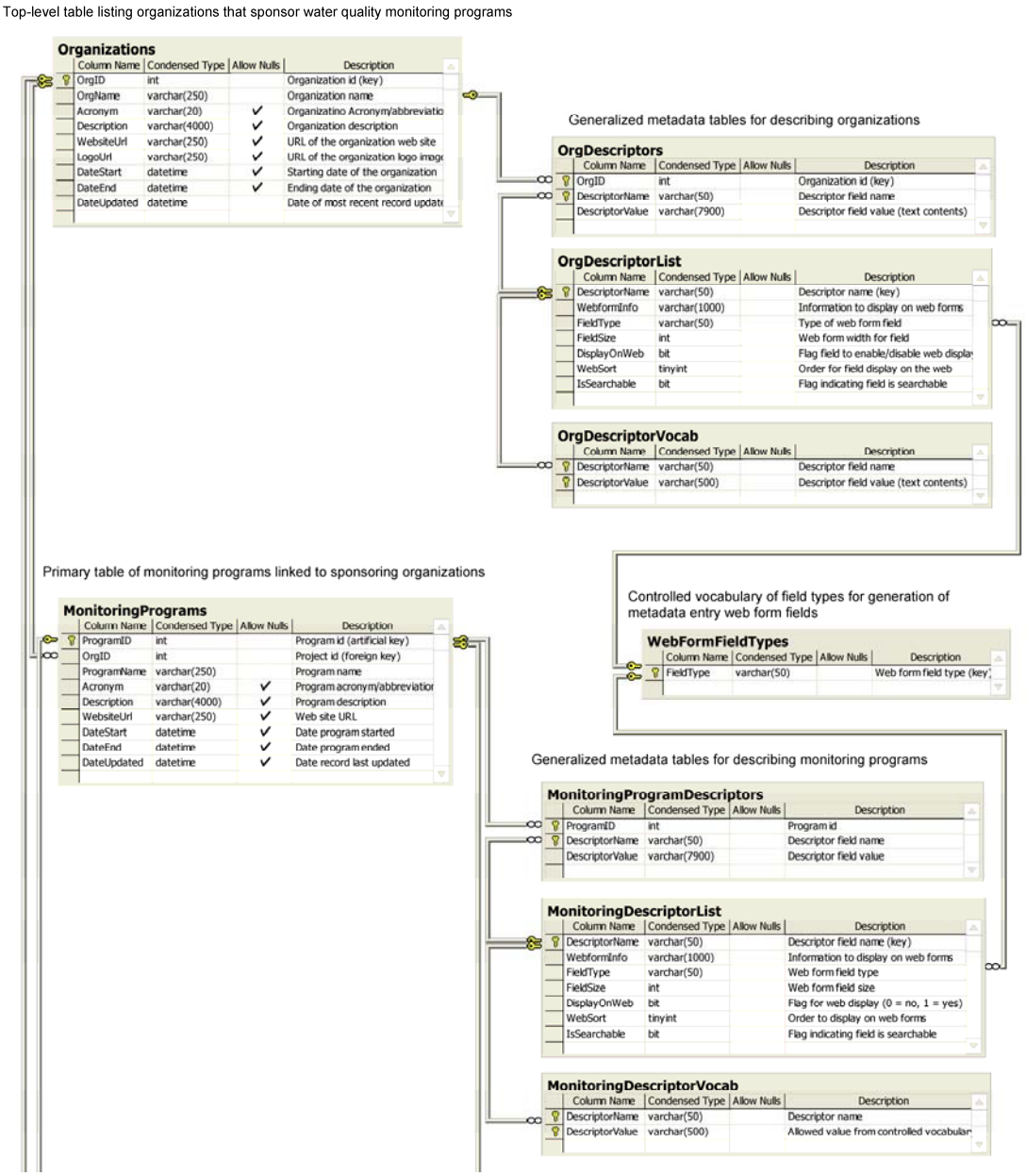

Figure 1. Entity-relationship diagram for organization and monitoring program elements of the NPS_WaterQuality database (version 1.0). Note that lines represent relationships between tables, with cardinality indicated by a key symbol for the primary table key and infinity symbol for the foreign table k

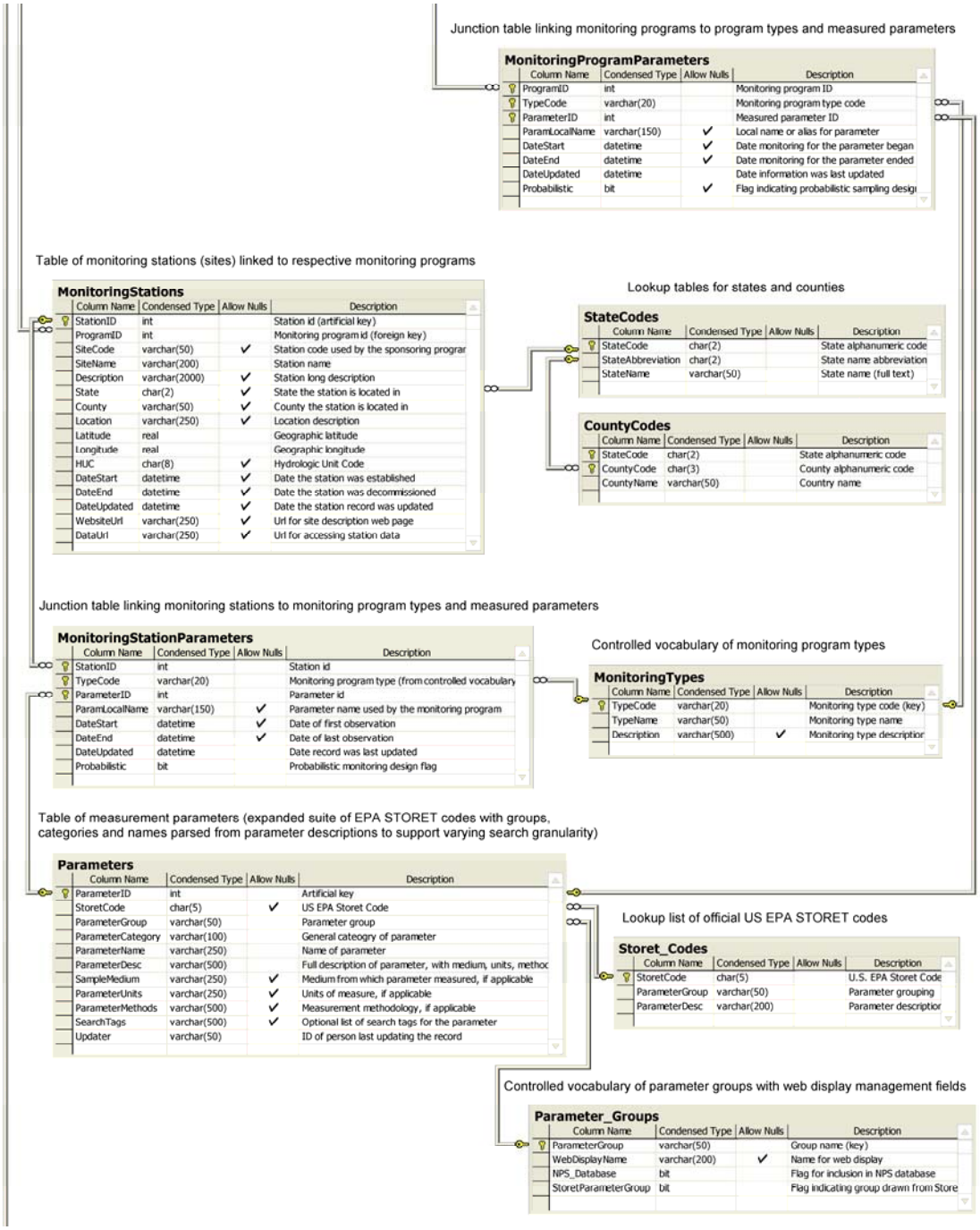

Figure 2. Entity-relationship diagram for monitoring location and parameter elements of the NPS_WaterQuality database (version 1.0). Note that lines represent relationships between tables, with cardinality indicated by a key symbol for the primary table key and infinity symbol for the foreign table key.

address, starting date, ending date, record update date). Additional tables are then linked to each of these primary tables (OrgDescriptors and MonitoringProgramDescriptors) to store varying amounts of metadata for a list of candidate metadata fields defined for each primary table (OrgDescriptorList and MonitoringDescriptorList). This generalized design allows individual metadata fields to be added, removed or deprecated without any changes to the underlying database schema or applications.

After designing the NPS_WaterQuality database, we used Microsoft Access™ 2003 to develop data entry and

management forms, standard queries (i.e. database views), and stored procedures to support web applications, web services and generate reports (Chipman and Baron, 2000). The forms and reports are stored in an Access Data Project (ADP) file along with database connection information, while the database objects (tables, views, stored procedures, functions, diagrams) are stored on the centralized server. This client-server approach allows multiple users to work with the NPS_WaterQuality database simultaneously without causing version-control issues or risking database file corruption, as commonly occurs with file-based databases.

Data Entry

GCRC staff entered information for organizations whose coastal water quality monitoring programs met our selection criteria (geographically located along the SE coast between the northern border of NC and the southern tip of Florida's east coast; occurring within the Coastal Zone demarcation for each state; having an existing or planned long-term data collection program with existing or planned data access). Metadata content was manually extracted from program web sites (e.g. program titles, descriptions, policy website links) or obtained from agency representatives. Program representatives were then asked to review the draft entries online to ensure accuracy. These initial efforts resulted in a total of 41 programs being registered in the database.

Where possible, we retrieved relevant monitoring location and parameter information from national databases (e.g. USGS National Water Information System, NERRS Centralized Data Management Office, LTER Data Catalog) by programmatically mining web applications and web services using custom MATLAB programs and uploading parsed information directly into the database. We requested information in spreadsheet form for programs that do not operate web-accessible information systems. Metadata for 16,182 stations with 286,042 station-parameter records were uploaded to the database through this initial effort. We also retrieved descriptions and geographic coordinates for over 15,000 additional aquatic and groundwater stations from the US EPA STORET database, and work is currently underway to extract location and parameter information from the EPA Water Quality Exchange database (WQX).

Database Querying

Despite efforts to obtain comparable information for each program, large disparities in the quantity of information available online and variations in reporting practices for associating measured parameters with locations complicated design of SQL queries for displaying and searching across the database. We therefore developed a series of standard queries, stored as "views" in the database, to create "virtual tables" that can be searched and displayed using simpler query syntax (Henderson, 2000). These views also provide a layer of abstraction for the underlying database structure, insulating dependent applications against future changes.

Stored procedures were then developed to support querying these database views using parameterized function calls to further simplify programmatic access (Henderson, 2002). In addition to encapsulating the query logic and improving performance, stored procedures also significantly enhance security in a web environment by removing database structure information from the application layer (Henderson, 2002). Each procedure is optimized for a different range of search parameters to maximize efficiency and performance, but all return the same fields in the query results to simplify formatting results.

Web and Map Interfaces

We developed a prototype web portal to provide broader access to the database. The portal is hosted on the Georgia Coastal Research Council web site (www.gcrc.uga.edu/wqmeta/) in the UGA Department of Marine Sciences. Web pages were designed using HTML templates and Active Server Page server-side scripting technology to dynamically retrieve information from the SQL Server database and format it for web display and downloading.

Web forms are available for entering and updating organization and program descriptions, along with additional pages that allow users to browse summaries and detailed metadata describing registered organizations and monitoring programs. Dynamically-generated hyperlinks support navigating to related information in the portal. One can also leave the portal to view external web links for the organization, program, or station (e.g. view the monitoring program's web site or access the data retrieval page).

A monitoring location search page allows visitors to search the database for monitoring activities by program, site name, site type, measured parameter, date range, 8-digit hydrologic unit code (HUC), and geographic bounding box (www.gcrc.uga.edu/wqmeta/app/search.asp). An embedded Google Map with zoom and drag capability is provided for graphical entry of bounding box coordinates, and measurement parameter names can be specific (e.g. Ammonium) or general (e.g. Nitrogen) to support broad or narrow searches. Multiple output formats are supported for search results, including HTML table (i.e. web display), Spreadsheet CSV, Google Earth Keyhole Markup Language (KML), and XML text. The HTML table format includes links to site detail pages, displaying all available information about monitoring activities at the site, links to monitoring program information, and an embedded Google Map. The KML format allows search results to be

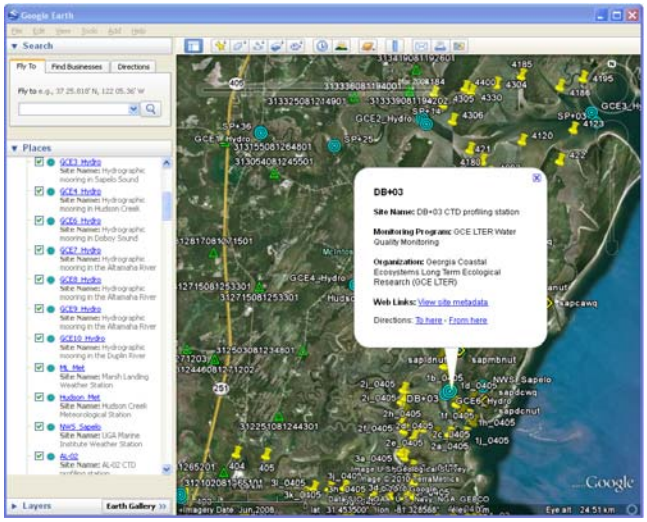

Figure 3. Query results in KML format displayed in the Google Earth application. Note that stations for each monitoring program are displayed using a different placemark, and metadata information and a link to the web portal is provided in popup balloons.

displayed in Google Earth and Google Maps. The KML implementation includes distinct placemark icons for each monitoring program, popup balloons with information and portal links for every site, and "TimeSpan" elements allowing the geographic coverage of stations to be displayed at varying points in time in the Google Earth application (fig.3).

In addition to the interactive web portal, web services were developed using Microsoft SQLXML 3.0 (Richardson and Ruby, 2007; Henderson, 2002). The web services allow sophisticated users or other portals to programmatically query the NPS_WaterQuality database over the Internet and retrieve information in structured XML or KML format. Descriptions of these services, including query parameters and output schemas, are available at www.gcrc.uga.edu/wqmeta/app/web_services.asp.

RESULTS AND DISCUSSION

Since it was first deployed in mid-2009, over 2000 web visitors have accessed the prototype web portal, accounting for over 12,000 page views. Use of the web services is not logged, so this is probably an under-estimate of database use.

In 2010, we received additional funding from NPS to continue identifying partners and adding new programs to the SE Coastal Water Quality Monitoring Metadata Database. We also extended the geographic scope to include the area defined as the Dept. of Interior's South Atlantic Landscape Conservation Cooperative (LCC), which ranges further inland than the initial area (fig.4). As part of this new project we have already added over 20,000 coastal monitoring locations by automatic harvesting of metadata from USGS and USEPA, and so will approximately dou-

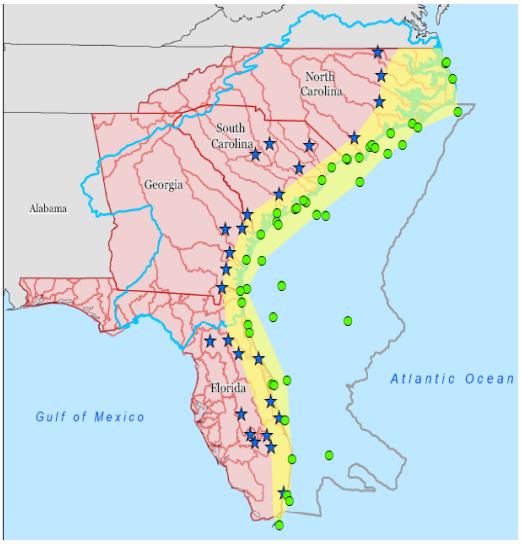

Figure 4. Geographic scope of the SE Coastal Water Quality Monitoring Metadata database, showing initial location (yellow shading); the boundary of DOI's Landscape Conservation Cooperative (blue line); and proposed seaward expansion to the Southeast Large Marine Ecosystem (grey line). Locations of NDBC buoys (green dots), USGS downstream gages (blue stars), and coastal HUCs (red lines) are also shown.

ble the number of location and measurement parameter entries in 2011.

As we continue to populate the database and coordinate with emerging regional coordination efforts (e.g. SECOORA and the NPS OpenParks GRID), we expect use of this resource to increase dramatically. Regional and neighboring networks and partnerships that have expressed support and interest in the practical outcomes of the project include Sea Grant's South Atlantic Regional Research project (SARRP), South Atlantic Marine Fisheries Council (SAFMC), Southeast Coastal Ocean Observing Regional Association (SECOORA), the Gulf of Mexico Alliance (GOMA) and Sea Grant's Mid-Atlantic Regional Research Project (MARRP).We are also pursuing additional funding to continue our efforts beyond the end of the current NPS project (June 2011), as well as improve the usability and functionality of the website to simplify end-user access to this information. In addition, we would like to begin analyzing the information we have compiled

to identify data gaps and help inform future sampling efforts.

CONCLUSION

The SE Coastal Water Quality Monitoring Metadata Database provides a set of valuable tools and several applications to a wide range of potential users. On a regional scale, this water quality database and the portal is potentially useful to a broad community of coastal managers, researchers, planners and others. This project provides a solid foundation for improved resource management by providing information that will help programs to prevent monitoring duplication and identify data gaps. We expect the portal to be a useful resource for regional coordination.

ACKNOWLEDGMENTS

We thank the participants of the June 2008 SE Coastal Water Quality Monitoring Metadata Workshop at Hollings Marine Laboratory for valuable comments on a prototype database design and web portal. Financial support was provided by the National Park Service (NPS Cooperative Agreement H5000 03 5040) and Georgia Coastal Ecosystems LTER Project (NSF Award OCE-0620959). The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NPS or NSF.

LITERATURE CITED

Chipman, M. and Baron, A. 2000. Microsoft Access Developer's Guide to SQL Server. Macmillan Computer Publishing, 864pp.

Environmental Data Standards Council (EDSC). 2006a. Environmental Sampling, Analysis and Results: Project [EX00002.1].

Environmental Data Standards Council (EDSC). 2006b. Environmental Sampling, Analysis and Results: Monitoring Location [EX000003.1].

Henderson, K. 2000. The Guru's Guide to Transact-SQL. Addison-Wesley Professional Press, 592pp.

Henderson, K. 2002. The Guru's Guide to SQL Server Stored Procedures, XML, and HTML. Addison-Wesley Professional Press, 800pp.

Richardson, L. and Ruby, S. 2007. RESTful Web Services. O'Reilly Media, 448pp.

Riordan, R. 1999. Designing Relational Database Systems. Microsoft Press, 293pp.

Sheldon, W.M. Jr., Laporte, C., Carpenter, J.P. and Alber, M. 2009. Southeast Coastal Water Quality Monitoring Metadata Tools - Database and Web Applications. In: Natural Resource Report NPS/SECN/NRR-2009/159. National Park Service, Fort Collins, Colorado.