# DATABASE ON SALINITY PATTERNS IN FLORIDA BAY*

N. Rishe, M. Chekmasov, M. Chekmasova, D. Hernandez, A. Roque, N. Terekhova, A. Zhyzhkevych
High Performance Database Research Center, School of Computer Science, Florida International University, University Park, Miami, FL 33199, hpdrc@cs.fiu.edu, http://hpdrc.cs.fiu.edu

## ABSTRACT

Salinity in Florida Bay is closely related to water management in South Florida. Water management activities over the last century have disrupted the quantity, quality, timing and distribution of freshwater flows into Florida Bay affecting salinity conditions. The main goal of the project presented in this paper is to accumulate all the data available on salinity in Florida Bay into one database and make this data available to the researchers and public via the Internet. This unified data source will give scientists one more tool to monitor the fragile ecosystem of the Everglades and to give better recommendations on water management in this area. The challenge of the project is in database design which will accumulate data collected by different groups of people who apply different methodology of data collection, different measuring equipment and techniques as well as different rules for data recording and formatting. Three major data sources on salinity conditions within Florida Bay are available. The three sources are historical data, temporal Everglades National Park (ENP) data, and spatial US Geological Survey (USGS) data. Historical data contains direct salinity observations from 1936 to present. ENP data include salinity and related parameters, such as rainfall, dissolved oxygen received from an increasing number of stations in the Bay continuously monitored by ENP since the early 1980's. The USGS dataset contains the spatially intensive bimonthly salinity survey records. Besides the aspects of the database design the paper covers implementation and maintenance issues. Also the web application is presented which provides access to the data via the Internet in a convenient and intuitive way without special knowledge of the database query tools.

## 1. SALINITY DATA AND REQUIREMENTS TO THE DATABASE

The salinity record for Florida Bay extends from the beginning of 20th century. The early records are not systematic and are usually found across a diverse literature and many unpublished sources. Despite the fact that these data sources are usually sparse, poorly formatted and are not present in the digital form, they compile a historical dataset which should in some form or another be reflected in the database. The historical dataset is of value. It has information on water salinity prior to the present rapid human development of South Florida that changes salinity patterns in Florida Bay, which in turn affects environmental conditions in the area.

In the past three decades with the extensive use of computers to store and process data, a number of salinity studies in Florida Bay resulted in a collection of the datasets which contain data in the digital form according to the well-documented formats. This data is ready to be stored in the database. However, the studies producing the data differ in methodology of collecting data and equipment used. Good example of the differences are NPS (National Park Service) hourly monitoring data, which represent time series from single locations, and USGS (United States Geological Survey) boat survey data which is spatial data with many locations represented by a single value at each location. This makes hard the work to unify the data in one database.

* Presented at the Seventh International Conference on Remote Sensing for Marine and Coastal Environments, Miami, Florida, 20-22 May 2002. This research was supported in part by NASA (under grants NAG5-9478, NAGW-4080, NAG5-5095, NAS5-97222, and NAG5-6830), NSF (CDA-9711582, IRI-9409661, HRD-9707076, and ANI-9876409), ONR (N00014-99-1-0952), and the Florida Space Grant Consortium.

Many salinity studies provide temperature data as well since the measurement equipment allowed to collect salinity and temperature data simultaneously. It was decided to store the acquired temperature data in the salinity database.

It is expected that users of the salinity database will be mostly interested in averaged salinity data over a particular area of Florida Bay and during a certain period of time rather than some particular salinity record. Daily, weekly, monthly, seasonal and annual averages are to be used. Since variations in data density are common from one study to another, data integration may lead to inaccurate results. In order to minimize this effect the following rules have been applied when calculating monthly, seasonal and annual averages: (i) these calculations are based on daily average data; (ii) daily average data are calculated as the average salinity/temperature for each station within each study on a given day, and (iii) in spatial data sets salinity/temperature observations are averaged within each basin on a given day and assigned a station location of the geographic center of the respective basin. We refer to (Robblee et al, 2000) for more information.

An extensive search has been performed by the researchers for literature references, published and unpublished, interpretable in terms of salinity conditions in Florida Bay. These references describe observations on salinity and other phenomena related to water quality like freshwater occurrences and fish kills. One of the requirements to the database was to facilitate storage and retrieval of these references.

## 2. CONCEPTUAL DATABASE SCHEMA

We have employed a semantic modeling approach for the database design. It has certain advantages in comparison with other techniques:

i. the output database schema design is intuitive and clear for understanding even by non-professionals in the field of databases
ii. the semantic schema reflects only the semantics of data to be stored in the database and does not show any technicalities of implementation
iii. semantic design can be automatically mapped to the relational schema on the implementation step, tools are available to produce instructions for physical database creation using any popular RDBMS (relational database management system).

Every concept defined on the semantic schema is one of the following:
a category - a set of objects about which information is to be aggregated in the database. Categories are denoted on schema by boxes with their names in uppercase bold inside the box;
an attribute - a pattern of certain printable data about the objects of a category. Attributes are denoted on schema by text in italic inside corresponding categories;
a relation - a pattern of relationships between objects of two categories. Relations are denoted by arrows on the schema. Certain other notations appear on the semantic schema like cardinality of the relations and indication of some constraints. We refer to (Rishe, 1992) for further reference on semantic schema designs.

We now turn to the description of the semantic schema of salinity database. For convenience purposes we are presenting the database schema comprising of three related subschemas, namely *studies*, *locations* and *datasets* subschemas.

A key concept of *studies* subschema is salinity STUDY (see Figure 1). STUDY is performed by a group of INVESIGATORs. One of INVESTIGATORs is a primary investigator, denoted by the relation **pi**. INVESTIGATORs are associated with INSTITUTIONs. Relation **works** links these two categories. Category REFERENCE is a container for published and unpublished papers and other documents related to salinity studies and possibly but not necessarily authored by INVESTIGATORs. This subschema accommodates the situations when some information on the reference is missed. For example, it allows to store information in the database in case, nowadays common, when investigators conduct studies and publish reports on the studies; as well as in case when historical reference is taken from a private diary never being published and not related to any study. Category KEYWORD contains keywords related to references and studies and is used to facilitate efficient search of

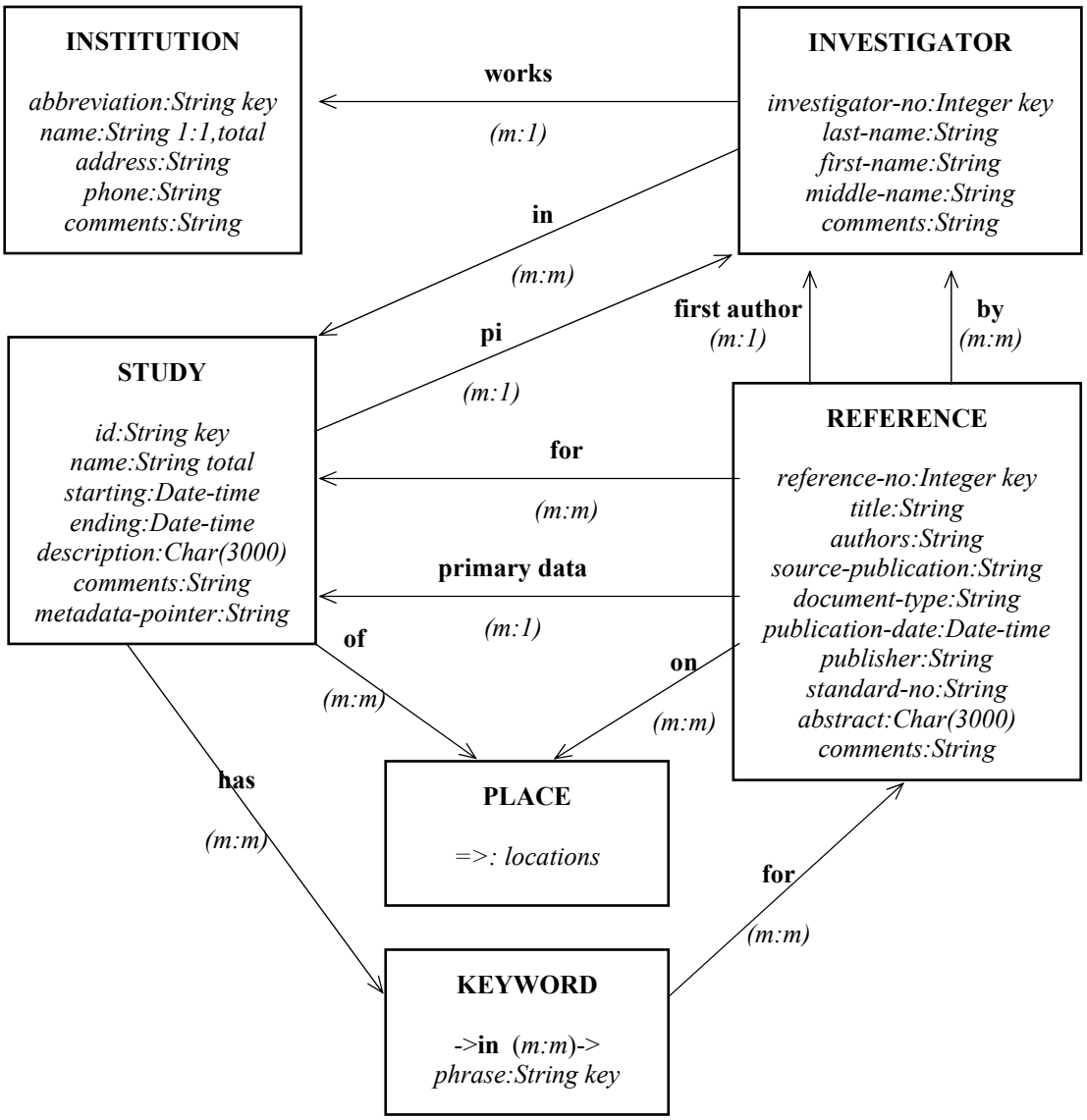


Figure 1. Semantic subschema *studies* for the salinity database.

references and studies by the user. Category PLACE, described below in subschema locations, has data on geographical areas, which can be mentioned by studies and references.

Subschema *locations* introduces a concept of OBSERVATION-POINT (see Figure 2). OBSERVATION-POINT describes geographical location where salinity measurements were made. It is characterized by latitude, longitude and a vertical position of the measuring instrument in water. Per user requirements the location information is duplicated in UTM coordinates, assuming that South Florida is in UTM zone 17. OBSERVATION-POINT can be part of the observation STATION and/or be located within BASIN. BASINs can be loosely defined as hydro-dynamically homogeneous areas. They form a grid, which completely covers the area of Florida Bay and the west coast river systems. On the design stage we assumed that while making queries to the database the user will be able to aggregate basins in order to define ecologically or hydrologically useful areas. PLACEs may vary in spatial extent and are related to basins. Thus PLACE names in the database have been defined in terms of a basin or

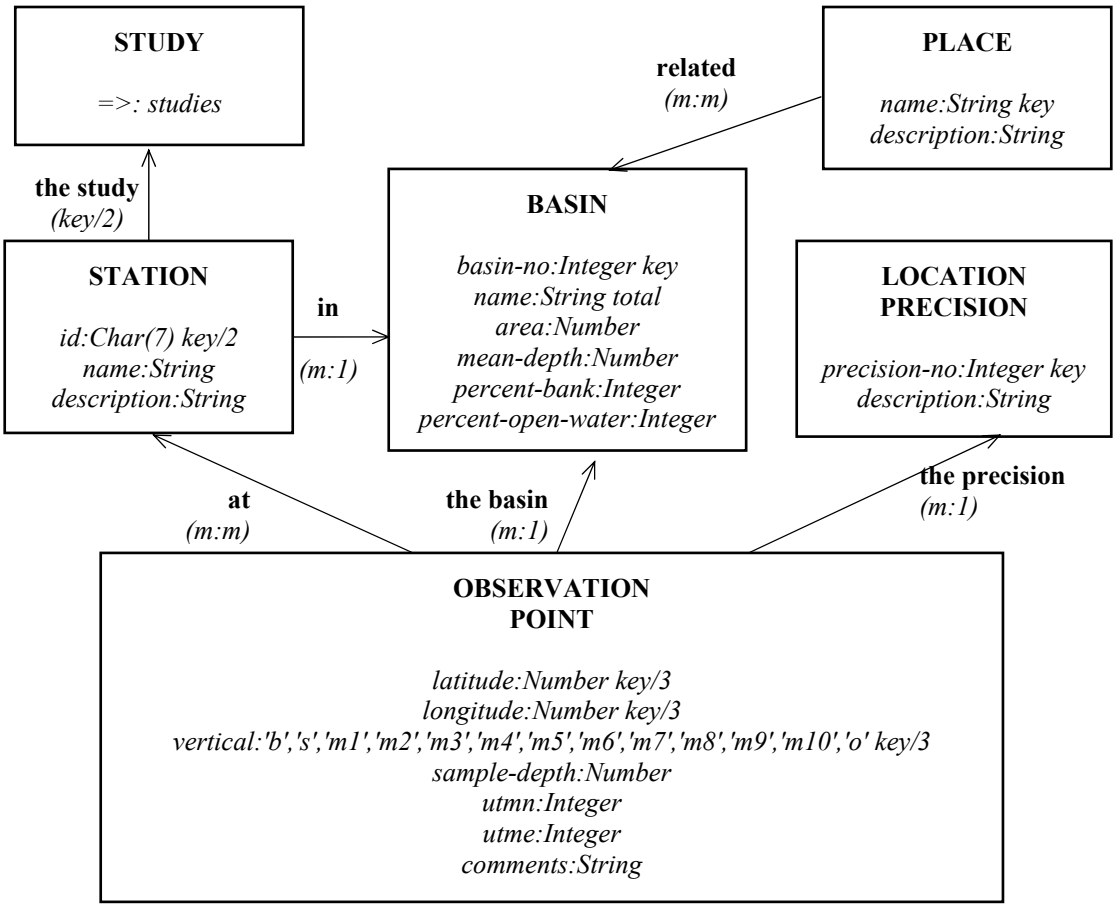


Figure 2. Semantic subschema *locations* for the salinity database.

basins. Since quite often we work with data collected from inaccurately recorded locations we had to introduce category LOCATION-PRECISION. In this category the information is stored on the confidence level of precision for some particular locations. OBSERVATION-POINT may also be part of a STATION. One station may have severalobservation points. At the same time we have cases in the salinity data when one station is identified by several *id*s within the study. The database design covers both possibilities by invoking many-to-many relationship between these categories. Finally, stations are considered only in the framework of salinity STUDY. Category STUDY was described above.

Subschema *datasets* is constructed around the concept of OBSERVATION-DATASET (see Figure 3). Dataset is uniquely defined by its *id*. It has starting and ending dates describing the period during which the records comprising the dataset were collected. Type of the dataset is also an important characteristic, usually temporal or spatial data are in the datasets. OBSERVATION-DATASET is collected using some METHOD and it is prepared for a particular STUDY. Categories SALINITY and TEMPERATURE identify storage for actual salinity and temperature data. Each record is being placed within a particular observation dataset. Also each record was collected at some particular geographical location (OBSERVATION-POINT).

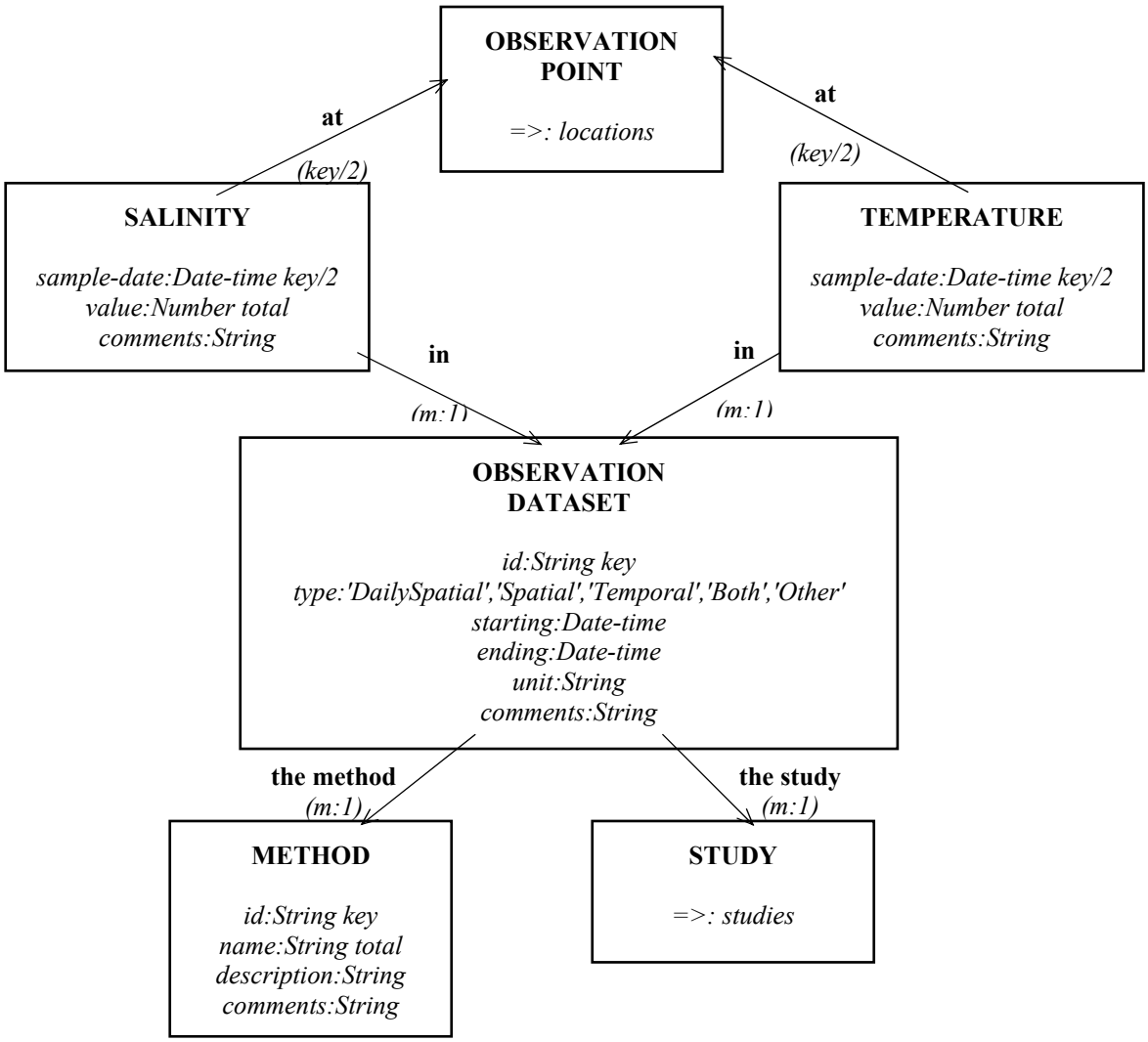


Figure 3. Semantic subschema *datesets* for the salinity database.

## 3. DATABASE IMPLEMENTATION

The salinity database is implemented using Oracle relational database management system (RDBMS). There are several major reasons to choose Oracle RDBMS: (i) It's one of the leading database management systems on the market; (ii) The clients already use Oracle for other projects and have a proper license for software; (iii) Oracle provides numerous tools for developing database applications, and in particular the Internet applications (iv) Oracle provides technical support, which proves to be helpful for database maintenance.

Based on the database design, the scripts were generated to physically create the salinity database tables and indexes, and to enforce the integrity and check constraints. The initial data was prepared and loaded using Oracle Loader tool.

Data loading further continues with the database being already in operational mode. It was decided to provide the users with two ways to load data. Large portions of data are loaded via Oracle Loader. However, several formats for the data files have been predefined and data files to be loaded have to conform to one of these formats. An example of a format could be plain text format with comma separated values having certain order of columns.

Another way to load data is manual loading via the Oracle forms. This approach is evidently slower but has certain advantages. The same forms for manual data entry may be used to do modifications of records already loaded to the database.

As we mentioned before in most cases the users of the salinity database are interested in averaged salinity and temperature data over a particular area of Florida Bay and during a certain period of time rather than raw data stored in the tables SALINITY and TEMPERATURE. To request that sort of aggregation using SQL (structured query language), a complex query statement should be written. We can not expect that sort of technical knowledge from the salinity database users. It was decided to provide scientists with a set of userviews, which are predefined, parameterized complex query statements stored in the database. The user calls the userview with specific parameters and gets the results.

With a set of raw salinity and temperature data steadily growing in the database, at some point in time the userviews provided will not meet the demands of the researchers to get aggregated data on the fly. When tables SALINITY and TEMPERATURE will be large enough, the processing of the userviews will take longer and thus will not be in real time. We are planning to solve this problem as follows.

The userviews collecting daily, monthly and annual averages over the whole salinity and temperature datasets will be run in advance and the results will be physically stored in the database as snapshots. These snapshots should be periodically updated to reflect all the possible changes in raw data, which may occur over a certain period of time. The easiest way to organize the snapshots' updates is to assign a periodic job to Oracle RDBMS (say, once a week) to recalculate the averages. Keeping in mind that these calculations will take considerable amount of computer resources, the time for the job should be chosen when the database is used less intensively, say during the nights. Having the averages for the raw datasets pre-cached in the snapshots will allow the system to significantly speed up response time to user requests related to averages, since these requests will be run against snapshots with most of the calculations already done.

## 4. WEB APPLICATION

We will now discuss the Internet application, which allows users to query salinity database and get the results via browser. This project is still under development so some functionality is not yet supported through the web. The benefits to provide the Internet access to the database are evident. The data (or part of it if some data is considered sensitive) will be available to general public and may be used to increase public awareness of the environmental projects done in Florida Bay. In particular, students may use the data for their science projects. The web access will allow the researchers to have access to data much easier. At the same time additional technical problems related to security and slow bandwidth via the Internet should be resolved for the web application.

When fully functional the application will assist users to start search for salinity and temperature data for particular basin(s) if the large areas are of interest. If the user focuses on specific geographical area, the data search may start by choosing place(s), particular observation station(s) or coordinates (latitude/longitude). It will be also possible to get data associated with some particular study(ies).

Figure 4 shows the query window of the web application for the salinity database when data search is conducted by basin(s). In this particular query we are concentrated in salinity data only. Thus the appropriate check box is clicked under **Dataset**. We are taking January 1999 as the time period we are interested in, the **Starting date** and **Ending date** are chosen on the form accordingly. We want to get daily averaged data and radio button in **Average term** is set on daily. We are ready to get data regardless of the water depth value where the measurement was taken: we click bottom, surface and intermediate check boxes under **Depth**. Finally, we are looking for the data from only one basin, Johnson Key Basin, which is chosen in the select list under **Basin** on the query form.

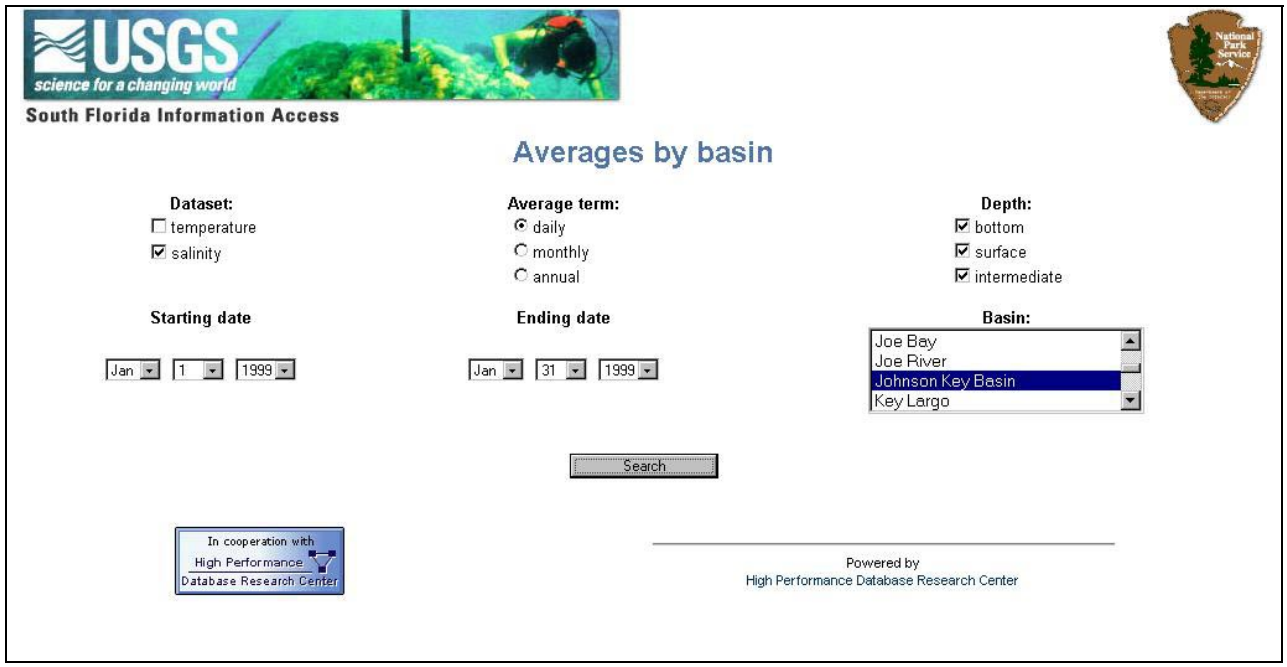


Figure 4. Query window of the web application for the salinity database.

Figure 5 displays the results for our query as a table with four columns. Column **Date** indicates the date when measurements were performed. Column **Depth** indicates the water depth for the measurements. Here 's' stands for surface, and 'b' stands for bottom measurements. Column **Parameter value** gives the daily averaged value for salinity measured in PSU (practical salinity unit). Column **Basin** gives the basin name, and has the only value in the case of our query.

Our sample query resulted only in few records, which can be easily analyzed on the screen of the Internet browser. However, in many cases the result set is expected to be thousands of records that need additional tools for further analysis. We are planning to provide two options for the user in this case. The first option is to download the result set as a file prepared in some commonly used format, say plain text with comma separated values. This file may be imported to any third party tool for further statistical analysis. This option will work as follows. The user will be asked to enter her e-mail address in the additional text box on the query form. When the user query is processed and the data file generated, the user receives e-mail with instructions how to download the file from the FTP (file transfer protocol) site, which supports the application.

The second option is to provide tools for data analysis within the web application. This includes changing column orders and providing sorting capabilities for the result set. Additionally generation of graphics to illustrate the data is planned to be implemented as an option. Finally we are planning to support functionality of pivot tables. Pivot tables are tools to work with data in multi-dimensional manner. For example, we can consider time as one dimension for salinity data, a set of station ids as another dimension and type of observation dataset as a third dimension. With the pivot table the user will be able to study the dependencies between the data elements according to the dimensions chosen. Pivot table will display data in a matrix-like form and will allow user to manipulate this matrix.

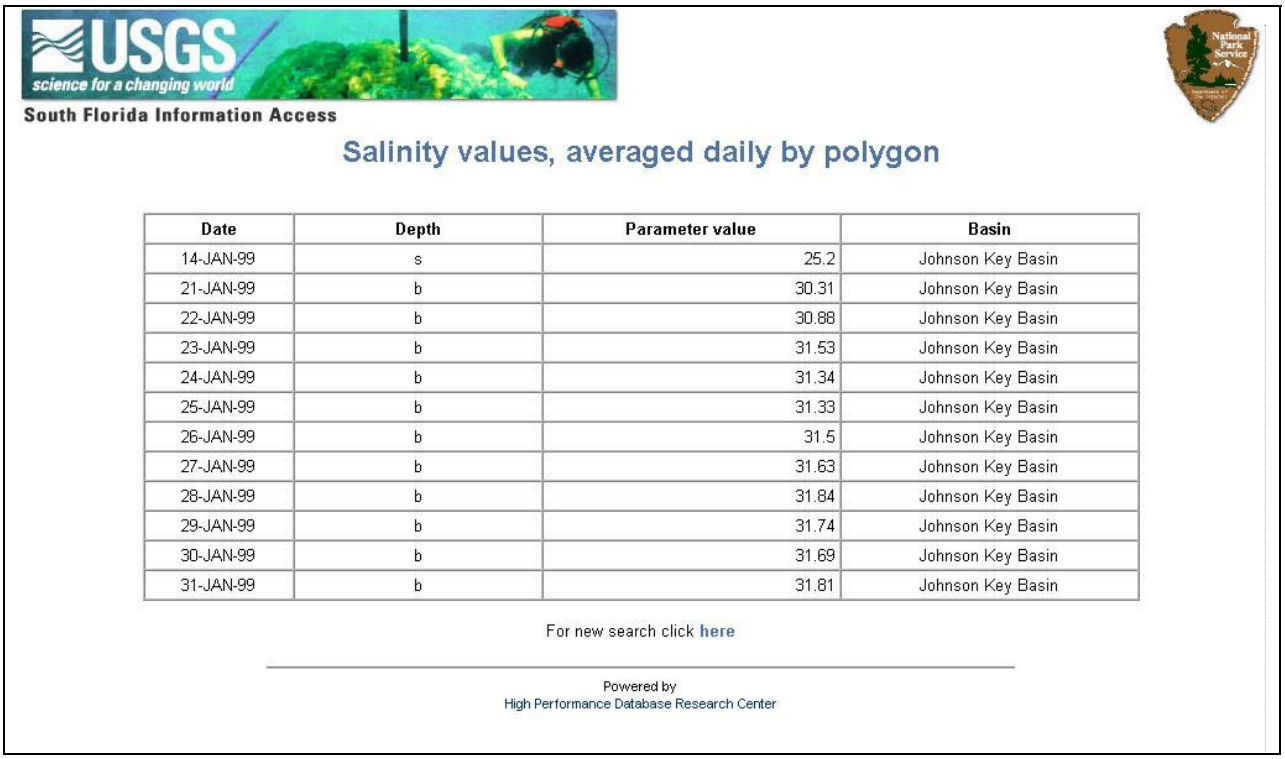


Figure 5. Query results window of the web application for the salinity database.

## 5. CONCLUSION

Design of the database to store salinity patterns in Florida Bay demonstrates a successful attempt to summarize all the data related to the subject into one well-organized and easily accessed data source. Implementation of the database design using Oracle RDBMS gives scientists an additional tool to study and analyze salinity data. Creating web application will broaden and further ease access to the salinity database. The framework of this project may be extended to include datasets of other environment studies performed in Florida Bay and generally in South Florida.

## 6. REFERENCES

N. Rishe, *Database Design: The Semantic Modeling Approach*, McGraw-Hill, 528 p., 1992.

M.Robblee, G.Clement, D.Smith, R.Halley, "Salinity Pattern in Florida Bay: A Synthesis", In *Proceedings of the Greater Everglades Ecosystem Restoration (GEER) Conference*, Naples, Florida, p.70-72, 11-15 December 2000.