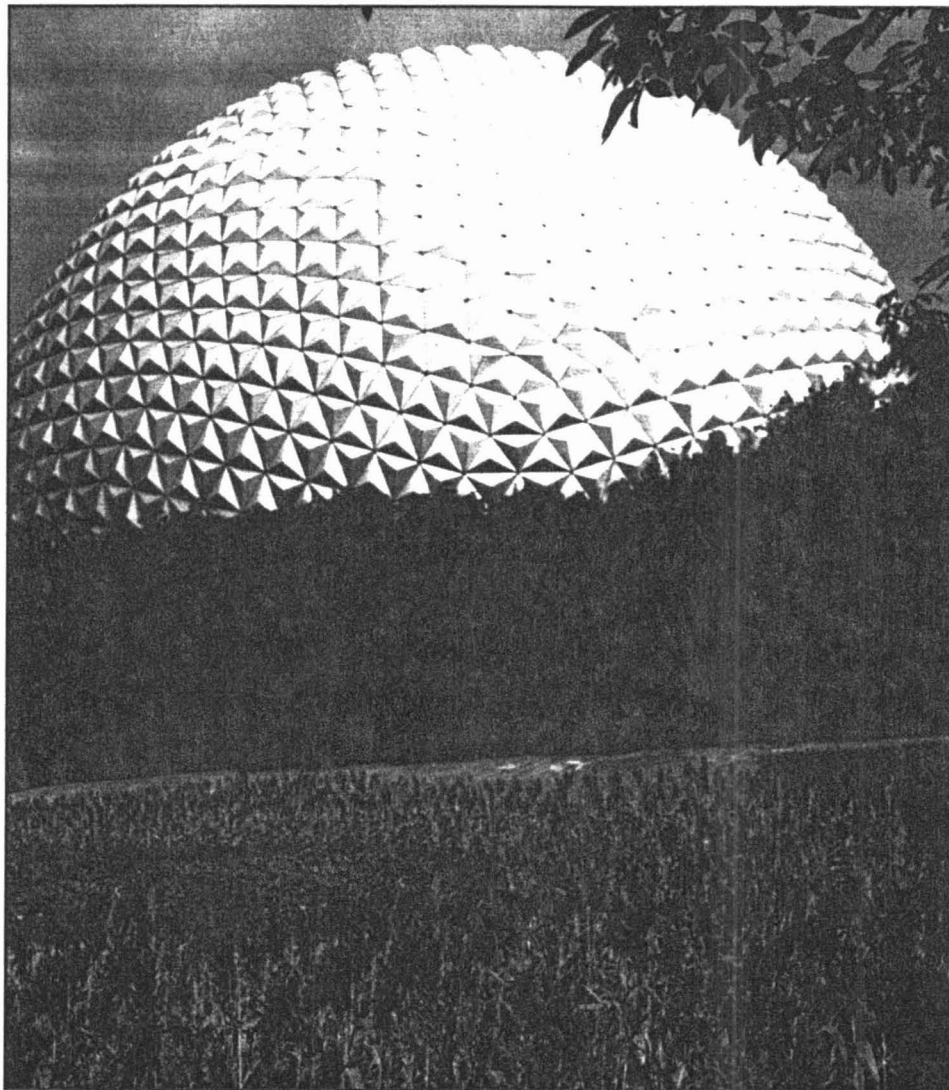*Proceedings of the First International Conference*

Geospatial Information in Agriculture and Forestry

Decision Support, Technology, and Applications

Volume I

1-3 June 1998
Lake Buena Vista, Florida, USA

SESSION E: DATA INTEGRATION FOR DECISION SUPPORT I

EVERGLADES DATA INTEGRATION USING A SEMANTIC DATABASE*

N.D. Rishe, D. Barton, M.V. Chekmasov, M.V. Chekmasova, S. Graham,
K. Madhyanapu, C. Pesantes
High Performance Database Research Center, School of Computer Science
Florida International University, University Park, Miami, FL 33199 U.S.A.
hpdrc@fiu.edu, http://hpdrc.cs.fiu.edu

ABSTRACT

The NASA Regional Applications Center at Florida International University's (FIU) High Performance Database Research Center (HPDRC) is involved to the acquisition, storage, and dissemination of all usable remotely-sensed and other data collected about the Everglades National Park (ENP) since the 1920's, as well as data collected for ground truth verification of current satellite observations. The purpose of this effort is to increase the use of remotely sensed information for monitoring the status of the ENP and the contiguous region. This paper describes a prototype of an environmental database that is being constructed to include all data related to the South Florida region. This data will be stored in a random-access semantic database for immediate use via the Internet. Methods already developed at HPDRC will perform the various data processing activities including data fusion, ingestion, compression, archived data retrieval, querying and mining, data publishing and dissemination, image rendering and visualization.

1. INTRODUCTION

There are several major reasons why the Everglades National Park attracts researchers involved in environmental studies. It not only has a unique and picturesque landscape, it is home to many endangered species. This area is highly affected by the human agricultural activities in the surrounding region and rapid urban development in South Florida faces problems such as supplying fresh water for the growing population. Thus, the Park indicates the environmental health of the whole area.

The overall project on constructing the Everglades Environmental Database (EEDB) is obviously a huge undertaking. This data came from different sources such as ground-based observations, aerial photography, satellite remote sensing together with demographic and cartographic information. In addition, the data is present in different textual and graphical formats. Even part of the project's goal of collecting the spatial data that is available for the ENP and the conservation area south of Lake Okeechobee is, by itself, a difficult task.

All the data should be integrated in such a way that a comprehensive database is available to researchers in a convenient, useful and timely manner. To solve the problem of accurate storage and management of the data collected, we employ the semantic modeling approach to the design and implementation of the database. This technique has been developed at the HPDRC and has proven to be

more efficient than the widely used relational approach for storing and retrieving large binary objects such as satellite images.

There are several aspects of systems research arising from this project aimed on extending the basic functionality of the database. Among the important goals are integrating of commercial Geographic Information Systems (GIS) with the EEDB and making the data available through the Internet.

The rest of the paper is organized as follows. Section 2 describes the types of data that will be stored in the database. Section 3 briefly describes the main principles of the semantic modeling approach. Section 4 is devoted to the programming tools and methods used for the design and implementation of the EEDB. Section 5 discusses the application of the database.

2. TYPES OF DATA FOR THE DATABASE

For the last seventy years the government and scientific organizations have performed a large number of environmental studies within the boundaries of the ENP. These studies have resulted in the collection of massive amounts of valuable data from various sources. The information from the following resources is to be populated in the EEDB: historical data (i.e., data collected about the Park since 1920's), ENP aerial survey data, satellite data of the terrain for the last 20 years, and data collected for the ground truth verification of the current satellite observations.

HPDRC is now completing the design and implementation of a large relational database for the Everglades National Park. This task is accomplished using Oracle relational database management system (RDBMS). More precisely, the relational database is a set of 22 databases, which store the results of environmental surveys performed within the Park during the past several decades. Some of these historical data sets will form an important part of the present semantic project and we now turn to a brief discussion of this data.

Data recording for a typical survey is performed as follows. The researchers have a set of printed forms, which they fill in with the data collected during fieldwork. In the office, the recorded information is entered into the computer and stored in the plain text files of a predefined format. This format is carefully described in the documentation. Nowadays the rapid development of computer technology allows using laptop computers during field observations for direct recording of measurements.

The files resulting from one survey logically form a data set. The largest data set stores the physical parameters of the area. These parameters include water depth, salinity, pH, temperature, wind direction and speed, air temperature, rainfall, conductivity, evaporation as well as others. A set of monitoring stations has been established within the Park to automatically record the physical data on a 10-minute, hourly or daily basis. All this data is populated into the Physical database. This data set is considered to be the most comprehensive and complete of the data sets available.

The other data sets resulted from vegetation and animal species studies. The Creel Census data set contains data on the fishing activities within the boundaries of the Park. Florida Bay, in particular, is an essential part of the Park and a vast region for exploration. The Shrimp data set focuses on the investigation of the life of invertebrates, but also contains data on marine vegetation observations, specific methods and stations. The aerial manatee surveys are found in the Manatee data set and a separate data set contains data on the study of Florida panthers' behavior. Several aerial surveys were conducted to study alligators' nests, wading birds and white tail deer location and activities. These studies formed the

Systematic Reconnaissance Flight data set. Smaller data sets contain information on Fire Ecology, Pine Demographic, East Everglades Exotic, Mitigation and other surveys.

Another source of information for the EEDB will arise from a joint effort between the Everglades National Park and the University of Georgia. This project is aimed at providing aerial photography of the Park region. This will make available a set of high-resolution images, which completely cover the Park's territory. These images are planned to be primarily used for geographic location, topographic and vegetation study purposes. The images can be scanned into the computer to be placed into the database.

Presently, satellite images are a valuable source of information that is used in decision making. Thus, this data should be populated in the semantic database. At present a low-volume satellite dish is installed at the FIU campus. It will be initially used to receive geostationary/geosynchronous operational environmental satellite (GOES) data for the EEDB. In the process of scientific exchange with other governmental and educational institutions, we are planning to receive images from other satellites. However some data still has to be acquired on the open market.

The variety of satellite images available is quite large depending on the characteristics of the corresponding remote sensing systems. The characteristics of these systems can vary in a number of ways. One way they can vary is in terms of spectral resolution. This refers to the number and dimension of specific wavelength intervals in the electromagnetic spectrum to which a sensor is sensitive. They can also vary in terms of spatial resolution. This is a measure of the smallest angular or linear separation between two objects that can be resolved by the sensor. Temporal resolution of a sensor can also vary. This refers to how often it records imagery of a particular area. Finally, radiometric resolution can vary. This defines the sensitivity of a detector to differences in signal strength. To highlight specific features of the terrain, filters may be applied to the image. This usually results in a set of supplementary images to be stored in the database in addition to the original one.

Other source of information to be populated in the EEDB is ground-truth verification of current satellite observations. The goal of these investigations is to verify whether conclusions made about biophysical variables that are based on the study of satellite images are true or false. This data can be obtained by comparing data collected by biologists in the field with data from the satellite images.

In addition to environmental data, related social-economic, demographic, transportation, and health information can be stored in the present semantic database.

3. SEMANTIC MODELING APPROACH TO THE DATABASE DESIGN

Widely used relational databases have provided good service in many conventional database applications. However, in situations where the structure of information is complex, greater flexibility is required (e.g., where objects with unknown identifiers or objects moving from one class to another are involved) or where non-conventional data is involved (long texts, images, etc.), other approaches such as semantic and/or object-oriented databases need to be considered.

The central notion of semantic database models is the concept of the object. An object is any real world entity that we wish to store information about in the database. The objects are categorized into classes according to their common properties. These classes, called categories, need not be disjoint - that is, one object may belong to several of them. Further, an arbitrary structure of subcategories and supercategories can be defined.

The representation of the objects in the computer is invisible to the user, who perceives the objects as real-world entities. These entities can be tangible, such as persons or cars, or intangible, such as observations, meetings, or desires. The database is perceived by its user as a set of facts about objects. There are three types of facts: (1) facts stating that an object belongs to a category; (2) facts stating that there is a relationship between objects; and (3) facts relating objects to data such as numbers, texts, dates, images, and tabulated or analytical functions. Relationships in the database can be of arbitrary kinds. For example, stating that there is a many-to-many relation 'address' between the category of person and text means that one person may have one address, several addresses or no address at all. We refer to (Rishe, 1992) for a discussion of the semantic modeling approach.

To illustrate some basic principles of the design of a semantic database, one can look at a schema design of a simple semantic database such as the schema found in Figure 1. This schema is designed to store the results of the so-called East Everglades Exotic Plant Treatment Survey.

Eight categories are present in the schema diagram, represented by rectangles. Among them three categories, namely **PROJECT, PERSONNEL** and **PROJECT-SPECIES-CODE,** belong to other semantic schemas and are referenced here. The rest of the categories have attributes which are written inside the rectangles. Arrows show relations between the categories. The direction of the arrow is from the domain to the range. The relations have also their own names.

The schema diagram preserves the logic of the Exotic Plant Treatment. One can read the schema to follow the logic of the survey. The main category of the schema is **SITE-TREATMENT**. It contains information on the treatment conditions of each particular site. Through **the-site** relation we are able to know the **site-number**, which uniquely identifies each site and its coordinates. By following **the-herbicide** relation we can find out which herbicide was used in the **SITE-TREATMENT, the-hydrological-conditions** relation link to corresponding information, and **the-species** relation tells which vegetation species were treated during the session. The treatment was done during **the-excursion,** which is uniquely defined by date and contains flight and labor information. In turn, the excursion (the category **EAST-EVERGLADES-EXOTIC-PLANT-TRATMENT-EXCURSION)** was conducted within **the-project**. Information on the Park employees who took part in the excursion and sites' treatment can be found by following **the-crew** relation.

A clear presentation of the database's business logic is one of the key advantages of the semantic approach to the design. A professional in any field is able to create a database design for his research without having to be an expert in the theory of databases.

4. ALGORITHMS AND PROGRAMS OF IMPLEMENTATION

A set of programming tools that can be used to design and create the semantic database, populate the data, and make queries and access the database through the Web is being created at the HPDRC. This technology has been implemented within the present project. In this section, we will briefly introduce the programs and libraries that are currently available.

A programming realization of the Semantic Database Modeling (SDBM) conception is the Semantic Database Interface (SDBI). The SDBI is a library of classes, functions and procedures, which are used for creating semantic database-oriented applications. Logically all the functions of the SDBI are divided into several groups. The first group of functions is responsible for creating, opening and closing
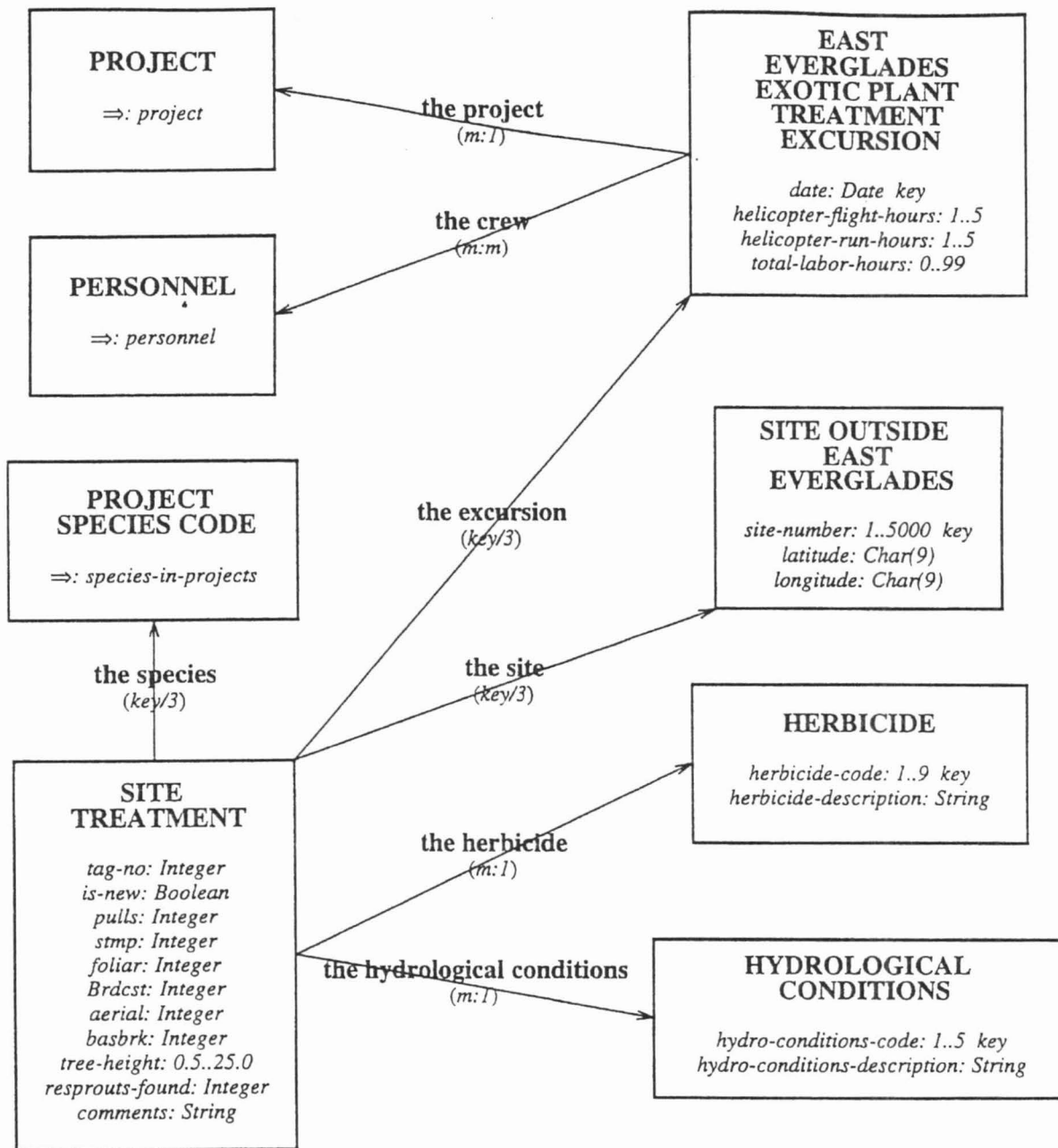
Figure 1. A Sub-schema For East Everglades Exotic Plant Treatment.

PROJECT

⇒: project

PERSONNEL

⇒: personnel

PROJECT
SPECIES CODE

⇒: species-in-projects

SITE
TREATMENT

tag-no: Integer
is-new: Boolean
pulls: Integer
stmp: Integer
foliar: Integer
Brdcst: Integer
aerial: Integer
basbrk: Integer
tree-height: 0.5..25.0
resprouts-found: Integer
comments: String

EAST
EVERGLADES
EXOTIC PLANT
TREATMENT
EXCURSION

date: Date key
helicopter-flight-hours: 1..5
helicopter-run-hours: 1..5
total-labor-hours: 0..99

SITE OUTSIDE
EAST
EVERGLADES

site-number: 1..5000 key
latitude: Char(9)
longitude: Char(9)

HERBICIDE

herbicide-code: 1..9 key
herbicide-description: String

HYDROLOGICAL
CONDITIONS

hydro-conditions-code: 1..5 key
hydro-conditions-description: String

the project (m:1)

the crew (m:m)

the excursion (key/3)

the species (key/3)

the site (key/3)

the herbicide (m:1)

the hydrological conditions (m:1)

the semantic database. When the database is opened, the updates to the database are performed within separate transactions. The transaction mechanism checks the database integrity, resolves any conflicts resulting from updates by several users and performs the physical update of the database. Another group of functions supports creating and editing the database schema, allows the implementation of user data types and works with categories and relations. Once a logical schema is designed and the corresponding database is created, the data loading proceeds. As was mentioned in Section 3, loading refers to

populating the database with facts about objects. A set of SDBI functions helps to complete the data loading process. Performing queries is a frequent operation while working with the database and several functions are implemented in the SDBI to support user queries.

Based on the SDBI, a set of applications has been created. These programs are aimed at automating the development process of the semantic database. The user is supposed to be concentrated on the semantic design. All the routine operations on creating the database are performed by the above mentioned applications. Some of these tools are described below.

One tool that is available is the Schema Editor. This is a Java application, which provides a user-friendly environment to create and modify the semantic schema design. By means of windows and dialogue boxes, a user is able to specify categories and relations, fill in comments, and use a mouse to locate categories and relations in the design. The final schema design is saved to a file in a special Schema Design Language (SDL) format.

The Schema Documentation tool is used to prepare the SDL files for printing. This tool transforms an SDL file to a Rich Text Format (RTF) file. The file can then easily be printed from MS Word or any other package, which supports RTF. Figure 1 shows a sample printout.

To physically create the database, the Semantic Database Generator is used. This tool takes an SDL-file containing the database design and produces a Semantic Database (SDB) file.

The Semantic Forms tool is currently under development. It is a Java program, which analyzes the structure of the database and provides the user with a set of forms that can be used to enter the data. It ensures that all the data to be entered meets the integrity and other constraints of the database.

For making queries, we have adapted Structured Query Language (SQL), the standard relational database language, to semantic databases. The original purpose of this adaptation was to be compatible with, and to be able to communicate with, relational tools. The application of SQL to semantic databases allows utilization of the full semantics of the data, can be applied to scientific and spatial data, properly treats missing values and produces queries which are typically an order of magnitude shorter than if written in SQL for an equivalent normalized relational database.

Several Semantic SQL related tools have also been developed. The Open Database Connectivity (ODBC) SQL server allows interoperability with third party tools. Embedded SQL allows programming of complex applications. A plain SQL server allows submission of ad-hoc and pre-programmed SQL queries on Windows NT and Unix machines. An internet/intranet World Wide Web (WWW) server allows submission of SQL queries via standard Internet browsers. An Internet SQL query generator helps the user to define a query without writing SQL statements. Finally, a graphical view selector allows the user to select a small view of the database against which to pose queries.

It is well-known that respond time of a user's query and the disk space necessary for storing data are the most critical characteristics for database development. In the present project we are planning to store satellite images in the database. These complex images cause the database to grow dramatically in size. At our site, we are conducting research to determine which image processing and data compression algorithms are most effective. Because the loss any data in compression techniques is undesirable for the satellite images, we are considering loss-less compression techniques. On the other hand, data compression slows down the respond time of the query. Further, queries performed remotely through the

Internet prove to be very time-consuming. Thus, we are investigating and implementing various algorithms designed to optimize the image loading and displaying process from remote directories and speed up decompression.

We believe that remotely sensed data should not be analyzed in a vacuum without the benefit of other corresponding information. As a part of our project, integration of the EEDB with the GIS will be done. To achieve this goal we are developing the GIS interface for the EEDB.

Summarizing, the tools and algorithms, which have been developed, or are presently under development, will help create a fully functional prototype of the EEDB.

5. APPLICATION OF THE EEDB

It is expected that a wide range of users will be interested in the EEDB. One of the main goals of the current project is to make NASA satellite sensor readings available for the needs of South Florida. By being effectively stored and managed, these satellite images are becoming a valuable source of information. Researchers and specialists in biology, urban planning, road construction, water management, medicine, weather forecasting and many other areas can use the images to gain valuable information necessary in their profession. In addition, the textual part of the database will contain a description of the images as well as other alphanumeric information gained from previously conducted environmental studies. The integration of the EEDB with commercial GIS will make the database even more useful and attractive. Availability of the database through the Internet will significantly increase the number of potential users.

Initially, the scientists conducting the biological projects will use the data stored in the EEDB for making reports, calculating statistics and planning new environmental studies. They will also provide new data that will be stored in the database in the future.

The students and faculty at FIU will also use the EEDB. This database will encourage scientific research collaborations among FIU academic departments and will be used for information retrieval and library bibliographic instruction purposes. This will be particularly useful for the existing GIS campus group at FIU and will aid in meeting the increased demands for GIS research/teaching at FIU. The teachers at Dade county public schools will be using the EEDB for educational purposes. Students will be able to receive information for their school environmental projects by visiting the EEDB web site.

Basing on this experience, it will be possible to establish relations with government agencies and private commercial enterprises in South Florida interested in new sources of information. This effort will expand the practical application of NASA activities in exploring the Earth from space.

ACKNOWLEDGMENT

We would like to thank the biologists and computer specialists from the Daniel Beard Research Center, Everglades National Park for their valuable assistance in the present project.

REFERENCES

N.D. Rishe, *Database Design: The Semantic Modeling Approach*, McGraw-Hill, New York, NY, p. 510, 1992.