

ARBOR Data Schema and Management Tools

Mary Pietrowicz, M. Pauline Baker
*Visualization and Interactive Spaces Lab,
Pervasive Tech Labs at Indiana University,
Indiana University-Purdue University Indianapolis*

Abstract

The Lilly ARBOR project (Answers for Restoring the Bank Of the River) is an experiment in ecosystem restoration, serving both educational and research purposes. The study gathers data related to habitat restoration, including measurements of trees planted at the site, records of water level and quality in monitoring wells, surveys of birds, butterflies, and dragonflies, and digital photographs of flora and fauna at the site. This paper describes the data schema established for the project, along with the collection of data management tools constructed for administering the data archive. This tool set provides a foundation for subsequent visual and statistical analysis of the ecosystem evolution.

1 Background & Analysis

ARBOR provides a site for the study of riparian ecosystems, evaluation of environmental reforestation strategies, and long-term monitoring and analysis. In addition, it provides a live laboratory and outreach service for students, faculty, and the community. The ARBOR project site is located in an urban area along the White River in the White River State Park, from 10th St. to New York St. in Marion County, IN.

The site is divided into 8 area plots of about one acre each. Each area is either a control plot or a test plot for one of three commonly-used reforestation strategies. The reforestation strategies include the following:

- 1) Use of 3-gallon containerized plants, planted on 12-ft. centers, with mowing and spot herbicide.
- 2) Bare-root seedlings planted in random pattern, with mowing and herbicide.
- 3) Bare-root seedlings planted in rows, with weed-inhibiting mat around trees, mowing, and herbicide.
- 4) Control plots with mowing, but no plantings.

Each reforestation strategy had two independent test areas. Twelve naturally-occurring tree species were selected and used in each of the plots [1] [2].

The ARBOR project staff also constructed seventy-two monitoring well compartments which they used to measure the water quality and water levels over time. Some of the well compartments were equipped with piezometers (pressure sensors) and automatic level loggers.

The ARBOR research team conducts wildlife (bird, butterfly, dragonfly, reptile, flora/fauna), tree growth (planted and naturally recruiting), seed bank, water level, and water quality surveys at the site. At the time this data schema definition project was initiated, a significant amount of data had already been collected. Most of the survey data had been recorded on a standardized paper form, checked for accuracy, and then transferred onto an Excel spreadsheet template (tree surveys, water quality, some water level surveys, bird/butterfly/dragonfly surveys). Some of the data existed as a Microsoft Word or Acrobat document (reptile surveys). Some of the data collection was automated (via water level loggers and GPS equipment). Even the water level logger data, however, was manually downloaded to a laptop computer in the field and then imported into an Excel spreadsheet. Table 1 summarizes the survey types, their corresponding collection process, and the data storage formats.

Survey Type	Data Collection Process Summary	Data Storage Format
Bird	Visual observation + Paper form entry + Excel entry	Excel
Butterfly	Visual observation + Paper form entry + Excel entry	Excel, Word
Dragonfly	Visual Observation + Paper form entry + Excel entry	Excel
Tree	Visual Observation + Manual measurement + Paper form entry + Excel entry	Excel
Natural Recruit	GPS automated data collection + Data entry into field device + Data download to Windows machine + Data export to Excel format	Excel
Reptile	Audio recording and observation + Field notes + Incorporation of data into report	PDF
Seed Bank	Pot soil samples + Pictures of growth +	JPG
Water Level	Manual measurement + Paper form entry + Excel entry	Excel
Water Level Logger	Automated measurement + Manual download to laptop in field + Export to CSV	CSV
Water Quality	Manual measurement + Paper form entry + Excel entry	Excel

Table 1: ARBOR Survey Type, Data Collection Process, and Data Storage Format

The ARBOR project team also collected a rich pictorial history of the project people, survey processes, and equipment [1]. Some project data was recorded entirely in pictures, for example, flora/fauna surveys and panoramic images taken from known locations. The pictures are manually downloaded, renamed, and moved to a meaningful location on the project web site. The file name and location in the directory tree usually suggested the kind of data that the picture represented, and the date on the file usually indicated the download date.

Some project data is static, that is, measured or listed once, but used repeatedly in data analysis. This kind of data included tree and wildlife species lists, tree planting information, and well installation information. Usually, location information was stored in ArcView, and the remaining information was stored in Excel spreadsheets. Species identification information was distributed across multiple sources, such as Excel spreadsheets, websites, and paper guidebooks. Please refer to Table 2 for a summary of static data type and the corresponding storage format.

Static Data Type	Data Description	Storage Format
Tree Planting	Species, Location, Planting Method, Planting Date	Excel, ArcView
Tree Species List	Common Name, Scientific Name, Numeric ID	Excel, Word
Well Installation	Installation Date, Well Type, Location, Vault Cover Elevation, Distance from Vault Cover to Casing, Distance from Casing to Level Logger	Excel, ArcView
Bird Species Info	Common and/or scientific name of species, expected occurrence at the site or in the area, picture, distinguishing characteristics, sample bird calls	Excel, USGS web site, Guidebooks
Butterfly Species Info	Common and/or scientific name of species, expected occurrence at the site or in the area, picture, distinguishing characteristics	Excel, USGS web site, Guidebooks
Dragonfly Species Info	Common and/or scientific name of species, expected occurrence at the site or in the area, picture, distinguishing characteristics	Excel, USGS web site, Guidebooks
Reptile Species Info	Common and/or scientific name of species, expected occurrence at the site or in the area, reptile calls, picture, distinguishing characteristics	PDF

Table 2: Static Data Type, Description, and Storage Format

The ARBOR team stored data in Excel and ArcView because it helped them to derive secondary data and generate graphs, charts, and maps to show research results. Please see Table 3 for an overview of the kinds of secondary data and visuals generated.

Derived Data Description	Tools Used/Format
Tree location and species plot on aerial view map	ArcView
Tree survival plotted on aerial view map	ArcView
Well types and locations plotted on aerial view map	ArcView
Bird Abundance by class over time (bar chart)	Excel
Bird Species Diversity by class over time (bar chart)	Excel
Well Water Temperature over time (scatter plot)	Excel
Well Level and Discharge data together over time (x,y)	Excel
Number of trees having caliper diameter values in a given survey (bar chart)	Excel

Table 3: Sample Derived Data and Graphics

2 Problem Statement

The ARBOR project has a rich, cross-disciplinary data set with defined processes for data collection. Exploring the data, producing visualizations, or doing statistical analysis of the data in its original form (Excel, PDF, Word, and ArcView) was difficult if not impossible. It was especially difficult to ask exploratory questions, or to analyze project data across multiple disciplines. An example question: How does water chemistry affect dragonfly population or tree growth over time? Exploring this kind of question requires the ability to query the data, visualize the query results, and generate tools to automate data analysis.

To facilitate data analysis activities, we proposed transformation of the set of collected data into a form with the following characteristics:

- Queryable
- Consistent in naming across survey and data types
- Standardized in the use of keywords
- Consistent in identification of people, surveys, etc.
- No unnecessary duplication of data
- Web-enabled for online data collection, analysis, and administration (no more paper required)
- Platform-compatible with common programming or scripting languages (e.g., Java, C/C++, Perl)
- Platform-compatible with common portal or web interface packages (e.g., Zope, Lasso, PhpNuke, JSP, Servlet, etc.)

We also proposed development and deployment of tools that will help collect, analyze, and visualize the data. Some desirable characteristics of these tools include the following:

- User-friendly query construction services (that do not resemble SQL)
- User-friendly query result set presentation services or browsers
- Statistical analysis services
- Standard chart/graph tools
- Voice interface to query and results analysis tools
- Services for analyzing, annotating, or enhancing pictures

- Event/notification subsystem
- User profiling for individualized data monitoring, presentation, and notification.

3 Database Selection and Schema Design

3.1 Database Selection

We decided to use MySQL [3] because it had bindings to all of the languages and portal packages we needed, performed well, ran on multiple platforms, was easy to maintain, and was free. At this writing, we installed version 4.0.10-0.

We also set up an Apache web server equipped with phpMyAdmin [4]. This gave us a GUI interface for database browsing and administration.

3.2 Database Schema

We designed the database schema after examining the existing ARBOR survey data, understanding the data collection and analysis processes, and reflecting on the applications we planned to build in the future. We designed it for maximum flexibility, ease of maintenance, and software reuse/simplicity [5]. Conceptually, the database contained the following kinds of objects:

- People/Profile
- ARBOR Zones
- Map Objects/Location
- Generic Survey Information
- Specific Survey Data
- Static Species Information
- Static Map Objects
- Pictures

The “People” concept had to support the unique identification of a person, definition of multiple roles for the same person (e.g., one person can have multiple roles within the same project or multiple sets of job titles and contact information), and a user profile outlining the users’ expertise and preferences. We encapsulated address information in a separate table because we expected that objects other than people will have addresses in the future. We also standardized roles, professional titles, proficiencies, and organization types by enumerating these in separate tables and preloading them. This approach is easy to maintain (just add/modify/delete items from the appropriate table). It also simplified the design of interactive applications (e.g., all of the valid job title choices were in one table) and analysis applications.

The “ARBOR Zones” concept allowed us to associate items with the eight ARBOR restoration zones and two transect lines. Transect line, area information, and planting method data was preloaded into the database.

The “Map Object/Location” concept allowed us to locate all pictures, wells, trees, and survey data on a map via GPS coordinates. Furthermore, the schema enabled the simultaneous use of multiple coordinate systems. This concept also gave us a quick way to associate an ARBOR area with a given xyz coordinate.

The “Generic Survey Information” concept included all of the data associated with a survey, regardless of the type. The ARBOR group always collected the survey date, weather conditions, names of surveyors, and location information. We ensured that the schema would support data commonly measured by weather stations because the ARBOR project planned to deploy a weather station. Encapsulation of a “survey header” in the schema encouraged similar encapsulation in software written to access and represent surveys.

“Specific Survey Data” tables encapsulate data specific to tree, water level, water quality, bird, butterfly, dragonfly, and natural recruit surveys. We further encapsulated well state (i.e., whether the well is flooded, dry, accessible, etc.) because these conditions are noted in all surveys involving the wells, and because it encouraged modular software design. In some cases, the survey content was likely to change, so we defined survey data attribute tables and preloaded them with standard attribute names for ease of maintenance. For example, a standard list of tree survey attributes would include whether the tree appeared to be alive, a demise code if appropriate, the caliper diameter, diameter at breast height, whether the tree showed evidence of predation, whether pruning had been necessary, tree height, whether the tree needed a new tag, and freeform comments.

“Static Species Information” contained standardized lists of tree, bird, butterfly, dragonfly, reptile, and other flora/fauna information. It included common and scientific names, species type, and when appropriate, picture and audio references. We preloaded this information since the data loading and analysis services required it.

“Static Map Objects” contained data about wells, trees, or other fixed objects in the ARBOR environment. These records must be loaded into the database before any attempts to load survey data are made. Each well or tree had corresponding map object data to specify location.

“Pictures” were special kinds of map objects. They referenced location information so that pictures could be “plotted” on a map, specified the picture type, contained a pointer to the URL or on-disk location of the picture, and specified descriptive keywords or attributes so that applications could classify and process pictures by type. Picture types, attributes, and keywords were standardized and preloaded, though additional keywords and attributes may be added in real time.

4 Database Import Package

The database import package provided both a way to import existing project data and an easy import/update path for future ARBOR survey data.

Most of the existing ARBOR survey processes required users to record survey data on standardized paper forms (Excel spreadsheet) in the field. The lab director would then check the field data for accuracy, and a staff member would type the survey data into Excel.

We decided that the least-disruptive, fastest way to import survey data and static map object information was to export the existing data from Excel to CSV, and write parsers that could read CSV, and develop data model and database interface classes that could insert or update database records.

Concept Object	Corresponding Database Tables	Preloaded Tables
People/Profile	person, personality, contact_info, org_title_list, org_type_list, proficiency_list, proficiencies	org_title_list, org_type_list, proficiency_list
ARBOR Zones	transect_list, area_list, planting_method_list, timezone_list	transect_list, area_list, planting_method_list, timezone_list
Map Objects/Locations	map_objects, xyz_coordinates, map_object_type_list, xyz_points_area	map_object_type_list
Generic Survey Information	weather, rain_list, snow_list, fog_list, moon_list, visibility_list, survey_data, survey_type_list, data_collectors	rain_list, snow_list, fog_list, moon_list, visibility_list, survey_data_type_list
Specific Survey Data	tree_data, valid_tree_data_attrs, tree_demise_code_list, nr_data, valid_nr_data_attrs, butterfly_data, dragonfly_data, bird_data, well_state, water_chemistry_data, valid_chem_measurements, well_level_data, well_level_logger_data	valid_tree_data_attrs, tree_demise_code_list, valid_nr_data_attrs, valid_chem_measurements
Static Species Information	Species_list, species_type_list	species_list, species_type_list
Static Map Objects	trees, natural_recruits, wells_type_list, wells	trees, wells_type_list, wells
Pictures	pictures, valid_picture_types, valid_picture_attrs, picture_attrs, valid_picture_keywords, picture_keywords	valid_picture_types, valid_picture_attrs, valid_picture_keywords

Table 5: Mapping of Conceptual Object to Corresponding Database Tables

We made minimal modifications to the existing project spreadsheets to ensure consistency with the database schema and ensure consistent naming across the project. For example, all surveys must have the same header, and people should be identified in the same way across the entire project. We added a new spreadsheet for identifying and profiling people, since this did not exist. We also restricted data entry fields to accept only reasonable data values and types, and we locked all cells except valid data entry cells. Note that these spreadsheets are not intended to be a permanent solution, but a first step toward a database-enabled, service-based, interactive system for data collection, analysis, and visualization.

The spreadsheet parsers were designed for the parser class to be a removable front-end to a set of data model and database interface classes. An interactive system, web-based or otherwise, may replace the function of the parsers in the future, and the model/database interface classes may be reused.

To date, we have completed parsers for Static Tree Planting Information, Static Well Installation Information, Tree Survey Data, and Water Quality Survey Data.

We have also completed data model classes to support the “people,” “map object,” “survey,” “tree,” “tree data,” “water level data,” and “water quality data” concepts. Each concept class implements a “Database Aware” interface so that it is equipped to handle inserting records, updating records, determining the need to insert vs. update, and reading from the database.

Since the parsers load data from a spreadsheet, the spreadsheets must have “primary key” data to identify each record uniquely. For the people loader, the unique identifier is the uits login. For the tree loader and tree data loader, the unique identifier is the tree id. And, for the well loader, well water level data loader, and well water quality data loader, the unique identifier is the well id.

5 Conclusions

Migration of the ARBOR data to a database was required before we could explore any data analysis or visualization applications. By using the existing ARBOR data collection processes, we could focus on the task of migrating the data with minimal disruption to the ARBOR project. And, by providing standardized spreadsheets with corresponding CSV parsers, we were able to provide a repeatable process for importing existing and future ARBOR data.

We may want to consider adding the ability for survey data objects to export and import themselves as XML, if requirements indicate, and infrastructure to send and receive survey data records as wrapped XML messages.

6 References

1. The Lilly ARBOR project web site, <http://www.cees.iupui.edu/>
2. Tudesco, Leonore, and Lindsey, Greg, “White River Riparian Restoration Project,” <http://www.cees.iupui.edu/research/wrep/index.html>
3. The MySQL Web Site, <http://www.mysql.com>
4. The phpMyAdmin Web Site, <http://www.phpmyadmin.net>
5. The ARBOR DAM Database Schema, <http://picard.uits.iupui.edu:8675/arbtor>