A Spatial Data Workbench for Data Mining, Analyses, and Synthesis

John Vande Castle and Deana Pennington
University of New Mexico, Department of Biology
Long Term Ecological Research (LTER) - Network Office
Albuquerque, NM, 87131-1091 USA

and

Tony Fountain
University of California, San Diego
San Diego Supercomputer Center, MC 0505
9500 Gilman Drive
La Jolla CA 92093-0505 USA

and

Cherri Pancake
Northwest Alliance for Computational Science & Engineering
Oregon State University
218 CH2M-Hill Alumni Center
Corvallis OR 97331 USA

ABSTRACT

Information managers at ecological research sites grapple with the complexity of diverse and heterogeneous datasets. The effective management of large geospatial datasets requires extensive hardware, software, and human resources that are often beyond the capabilities of smaller institutions. A major challenge has been the lack of tools capable of integrating very large geospatial datasets with more conventional ecological data. A data "toolbox" was developed to address this challenge. The toolbox, known as the "Spatial Data Workbench," consists of multiple layers of software tools that make it possible to access and integrate multi-temporal and multi-site geospatial data. The Spatial Data Workbench provides access to the large-scale data acquired for individual projects and makes them available in a user-friendly environment. The goal is to reduce duplication of effort and extend access to a wider research audience to facilitate integrative types of research, such as time series and cross-site analyses.

Keywords: Geospatial data, remote sensing, hyper spectral data, informatics.

1. ECOLOGICAL DATA MANAGEMENT

Ecological research sites such as those within the Long Term Ecological Research Network Program (LTER) of the National Science Foundation (NSF) are interested in the management of diverse ecological datasets. Information managers at ecological research sites grapple with the complexity of diverse and heterogeneous datasets. The effective management of large geospatial datasets, such as those generated by remote sensing, requires extensive hardware, software, and human resources that are often beyond the capabilities of smaller institutions. Significant amounts of staff time must be dedicated to daily data collection and monitoring efforts. Important datasets, especially those in the multi-gigabyte size ranges such as hyperspatial or hyperspectral data, often must be stored off-line, so individual files must be moved back on-line when needed for analysis. The integration, analysis and synthesis of those datasets require significant levels of staff time and expertise. The capability and capacity of hardware and software systems are just beginning to permit the manipulation of these valuable datasets as online resources.

The Spatial Data Workbench is a collaborative effort between LTER (http://www.lternet.edu/) and the San Diego Supercomputer Center (SDSC), (http://www.sdsc.edu/) as part of the National Partnership for Advanced Computational Infrastructure (NPACI) Earth System Science initiative (http://www.npaci.edu/). Data are managed with the Storage Resource Broker (SRB) at SDSC. The SRB provides the basis for integrative analysis tools, for the storage and dissemination of geospatial LTER datasets through World Wide Web and other forms of access.

2. THE SPATIAL DATA WORKBENCH

A major challenge has been the lack of tools capable of integrating very large geospatial datasets with more conventional ecological data. Advanced remote sensing data such as AVIRIS (Airborne Visible Infrared Imaging Spectrometer) data (Figure 1) are particularly difficult due to their complex structure and data volume.

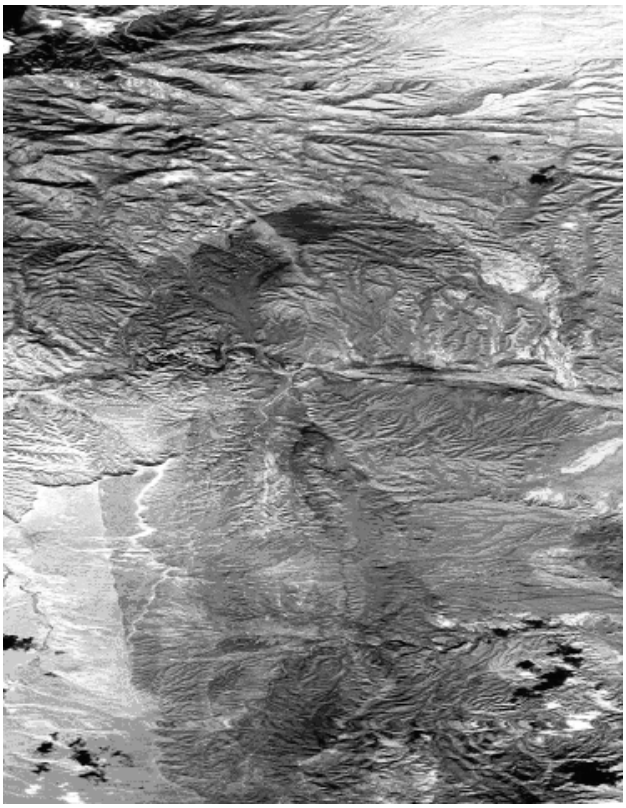

Figure 1. An AVIRIS thumbnail image representing the 224 bands of the full hyperspectral dataset.

The collaboration between LTER and SDSC developed a data "toolbox" to address this challenge. The toolbox, known as the "Spatial Data Workbench," consists of multiple layers of software tools that make it possible to access and integrate multi-temporal and multi-site geospatial data. In particular, the Spatial Data Workbench provides integrated support for large-scale data management and analysis using high-performance computing and storage facilities. It grew out of a need to manage and process the large AVIRIS hyperspectral datasets for several sites participating in the LTER program.

A single AVIRIS "acquisition" can encompass more than 50 gigabytes of data (Figure 2), typically distributed across a number of physical tapes. The desired datasets must be ordered from NASA's data archives, and then copied from tape to disk, a process itself requiring significant time and effort, before it can be analyzed. Once the analysis has been performed, the data are usually deleted because of the large data volume, to make room for other data.

Figure 2. An example of the data volume from a single 2-day data acquisition of AVIRIS data distributed on high capacity tape cartridges.

The capability of integrating hyperspectral data with more conventional remote sensing data such as Landsat Thematic Mapper was a secondary motivation for the project. The concept of the Spatial Data Workbench, then, is to make large-scale geospatial data more generally available, thereby reducing duplication of effort and extending access to a wider research audience. To do this, the Spatial Data Workbench is in part a data archive and access tool to preserve the large-scale data acquired

for individual projects. It makes the data available in a user-friendly environment that facilitates more integrative types of research, such as time series and cross-site analyses. More than an access to datasets, the design of the Spatial Data Workbench permits analysis in a high performance computer environment.

3. DATA ACCESS

The Spatial Data Workbench manages AVIRIS, Landsat Thematic Mapper, Advanced Very High Resolution Radiometer, and other remote sensing data. These data have been acquired by the Network Office of the LTER Program since 1990. Because the kinds of integrated analyses needed by ecologists require an ever-expanding number of tools and systems, our software toolkit consists of 3 tiers (client, server, and analysis) that provide the software "glue" for building plug-and-play systems (Figure 3).

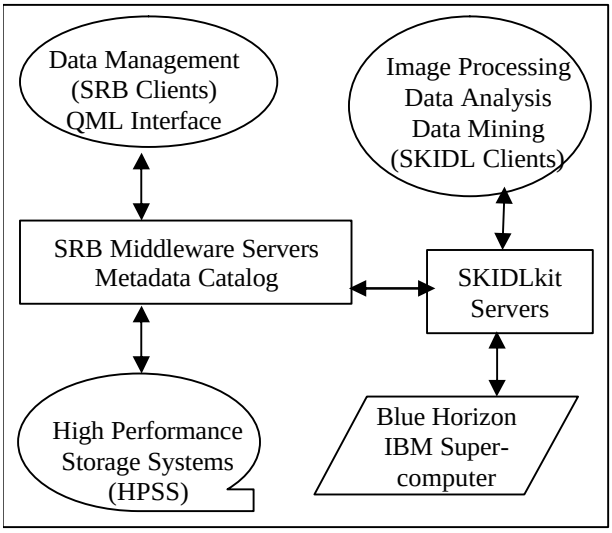

Figure 3. Architecture diagram and processing flow of the LTER Spatial Data Workbench.

Data residing within the Spatial Data Workbench are organized within a Storage Resource Broker (SRB), developed at the San Diego Supercomputer Center (SDSC). The SRB provides the middleware services needed to manage multiple, distributed and heterogeneous datasets as a single logical collection. The SRB provides seamless access to archival resources and file systems along with tools for authenticating users, controlling access, and auditing accesses. The SRB's metadata catalog, which includes both system-level and user-defined metadata, facilities query of individual datasets based on attributes rather than file names or physical locations. Metadata based on standards currently being developed specifically for ecological data (Ecological Metadata Language) by the NSF-funded Knowledge Network for Biocomplexity will be loaded into the SRB's metadata catalog. Advantages of the SRB include location transparency, improved reliability and availability [3]. The overall architecture is designed to be flexible in order to support a variety of configurations and applications.

Both metadata and datasets may be conveniently accessed through any web browser. The Spatial Data Workbench includes a variety of web-based search tools designed to reduce the level of technical expertise needed to locate remotely sensed data. The project web site, http://www.lternet.edu/technology/sdw/, includes documentation of available data cataloged by LTER site, links to data access interfaces, location information such as flight lines and imagery metadata, and low resolution browse images. An initial web-based interface was developed for direct access to the datasets although a more intuitive windows-based graphical user interface (Figure 4), to the Spatial Data Workbench is provided by a browser client to the data managed by the Storage Resource Broker at SDSC [1], [2]. We are also pursuing remote service access for the SKIDLkit tools by a Java multi-tier system and a second web services approach using XML, SOAP and WSDL. The services will be published so that developer can access them for custom applications, but the specific services will still need to be defined.

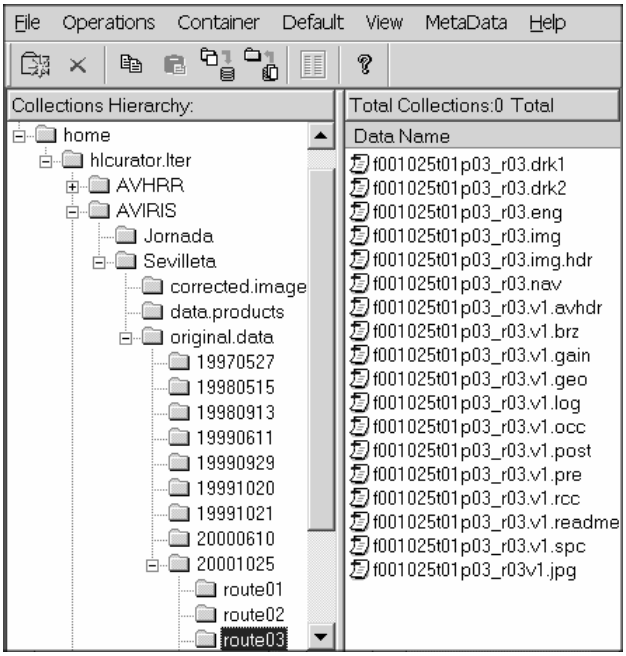

Figure 4. A windows-based browser interface to the Spatial Data Workbench data contents managed by the Storage Resource Broker at the San Diego Supercomputer Center.

We have also developed a specialized web-based interface to the high-volume, high-dimensionality AVIRIS hyperspectral data using Query Markup Language. The interface provides more simplified access to the more complex AVIRIS data sets. The custom interface is being extended to provide access to other imagery types.

4. DATA INTEGRATION AND PROCESSING

Due to the extreme size of these datasets, the Spatial Data Workbench must support server-side data manipulation and analysis. To augment the management capabilities provided by the SRB, we have developed a processing and analysis integration toolkit. SKIDLkit (Figure 3) is a Java-based client-server application with utilities for loading large databases, performing statistical/numerical analyses, and operating efficiently in a Grid computing environment. "Grid-based" in this context means computational grid, not a data matrix. We are tracking the Grid developments, e.g., the Grid Forum actions, and designing a compatible architecture. SKIDLkit provides the software glue for building applications, including interfaces to commercial database systems (e.g., DB2, Oracle), hyperspectral processing libraries (e.g., IDL), data mining packages (e.g., IntelligentMiner), and GIS and data visualization support (e.g., Polexis). Combining high-performance analysis tools with SRB-based collection management services enables scientists to explore and add value to their projects seamlessly and efficiently.

Processing and analysis tools are currently under development for exploring these data collections to discover patterns, create models, and test hypotheses relating to biological/ecological processes. A processing and analytical pipeline is being developed off-line for the AVIRIS hyperspectral data and is currently under test. This pipeline will be completed and integrated into the SKIDLkit interface providing server-side processing for many of the compute-intensive and data-intensive processing/analysis operations. The data will be automatically downloaded into a database, retaining spatial and temporal information. An SQL tool will be provided for database query, along with a selection of pre-written SQL statements for common queries. We are developing scalable implementations of algorithms for data reduction and feature selection in high-dimensional data sets using concepts from Bayesian networks, genetic algorithms and support vector machines. Prototypes are currently under test.

CONCLUSIONS

The Spatial Data Workbench was initially designed as a tool for managing very large volumes of geospatial data sets in an ecological research environment. The integration of the data into a high performance collection management system such as the Storage Resource Broker at the San Diego Super Computer Center provides the capability for distributed access and a pipeline to advanced analytic tools. The goal is to make large geospatial datasets more available in a user-friendly environment and to facilitate collaborative research.

This project uses tools including the Storage Resource Broker developed at the San Diego Supercomputer Center, University of California-San Diego. The supporting structure of the Spatial Data

Workbench, and software support to the Storage Resource Broker were made possible by the work of Arcot K. Rajasekar, of SDSC. This work is supported in part by a grant from the National Partnership for Advanced Computational Infrastructure (NPACI) Program: Earth Systems Sciences Thrust Award# 10152753 and by a grant from the National Science Foundation, #DEB-9634135 to the LTER Network Office.

REFERENCES

[1] NPACI, The SDSC Storage Resource Broker, 2002 - SRB web document at http://www.npaci.edu/DICE/SRB/

[2] Collection-based Persistent Archives, A. Rajasekar, R. Marciano, and R. Moore, *16th IEEE Symposium on Mass Storage Systems* , March 1999.

[3] Rajasekar, A.K. and M. Wan, 2002. SRB and SRBRack-Components of a Virtual Data Grid Architecture *Advanced Simulation Technologies Conference* (ASTC02), San Diego, April 15-17, 2002.