

# NatureServe Network Internet Data Delivery Project Update

August 2005



## **PROJECT APPROACH**

In May 2004 NatureServe was awarded a three-year NSF grant intended to improve online access to the biodiversity data resources held by NatureServe and its network of natural heritage data centers (DBI-03454000). To achieve this goal, NatureServe is establishing an Internet gateway for the distributed network of state biodiversity databases. Using an open, web services architecture for online access to these data resources, we will be able to deliver biodiversity information to the user more directly and effectively than ever before, in the format that the user needs, such as GIS data layers.

The Internet gateway is being built on two complementary technologies. The first is a suite of web services delivering both standard and custom biodiversity XML documents, discovered using UDDI and accessed using SOAP. The second is a set of robust GIS map services and feature services using the ESRI technology platform as its core. Both of these technologies have been combined so that they are platform and application independent to allow for the maximum user base possible. These two technologies form the data delivery portion of the Internet gateway.

Securing these two core technologies is a unique access control system that is being developed in collaboration with Bob Morris and a team of graduate student researchers at the University of Massachusetts-Boston. Most web service-based security approaches focus only on protecting XML documents, without addressing the unique problems that are presented by protecting GIS data integrated with other XML data. This robust security system is the key to reassuring data providers that their data will be protected and controlled. These technologies integrate to form a peer-to-peer network where all of the data providers in the NatureServe network are serving their data using a common set of web and map services. Using web services, NatureServe network data providers will eventually be capable of responding directly to information requests with near real-time data.

Project implementation is phased to allow initial access to network-wide geospatial data directly from the enterprise geodatabase, with a subsequent version allowing distributed queries of web-enabled local databases. The current NSF award will fund completion of the first phase, and will also provide funding for two local database pilot sites during the third year. As part of our efforts to develop the capacity of local

databases to respond directly to such queries, through our Delaware State partners, we have received an award under the EPA Environmental Information Exchange Network challenge grant program that will enable us to deploy web services in four participating state programs: Delaware, Washington, Illinois, and New Mexico. The web services technology developed through this challenge grant and tested in these four participating states will provide a foundation for replicating these services throughout the NatureServe network. We are optimistic that even wider deployment of web services at the local database level will be further enabled in future years through EPA Exchange Network implementation grants.

## **YEAR 1 ACTIVITIES AND ACCOMPLISHMENTS**

During the first year of this project, we were focused on understanding the requirements for providing secure access to distributed biodiversity data. In particular, we hosted a workshop attended by both data providers and data users to better understand data access requirements. We also conducted research into comparable on-line data delivery systems, including those of our own network members. Research on access control mechanisms was conducted by co-PI Bob Morris and his UMASS-Boston team. Beyond these research activities, a number of deliverables were also accomplished during the first year. These included:

- 1) Establishment of an enterprise geodatabase to serve as the spatial repository for NatureServe's element occurrence (EO) data;
- 2) Establishment of a DiGIR-compliant web service that provides online access to selected portions of the EO data that could be mapped to the Darwin Core standard, including county-level location information;
- 3) Case statements and user stories documenting requirements for the online system, from the perspective of NatureServe staff, data users, and data providers; and
- 4) Establishment of the necessary components for a web services delivery framework at NatureServe.

### *Data Access Requirements Workshop*

A data access requirements workshop was held in Tucson, AZ on November 17, 2004 and constituted one of the major planning and research activities for the project. This workshop was designed to present potential technology approaches to both local data contributors as well as key representatives of user communities to ensure that the system approach will meet requirements of both audiences. The workshop was held in conjunction with NatureServe's annual leadership conference, which allowed us to include a large number of participants. Approximately 40 individuals attended the workshop, representing at least 27 different institutions. These included state agency-based natural heritage programs, federal agencies, and academic researchers. Research carried out in our evaluation of existing systems was presented to workshop participants as a way of evaluating which of these approaches would be most suitable for deployment in this project. In particular, our Colorado State University partners developed a matrix characterizing functionality of different systems with respect to levels of data access and data resolution. In addition, we presented the results of analyses based on current data

sharing agreements that have been negotiated and executed between NatureServe and its collaborating natural heritage programs, which identifies levels of data resolution that might be available to different audience segments with varying access rights. Three breakout groups addressed the following issues in detail: criteria for user authorization and authentication; refinements to the data access matrix; and sustainability models. Key findings from the data access requirements workshop include:

- The biggest issues surrounding deployment of the distributed data delivery system are social in nature.
- The distinction between web services and traditional web environments is not well understood by data providers. Clarifying this should allow data providers to understand that many of their concerns can be addressed technologically. The development and documentation of use and “mis-use” cases will further that understanding.
- Discussions revealed a general validation for unrestricted access to USGS quad-level resolution data (an approximately 10km x 10km grid). However, additional discussion will be required to determine the specific attributes and functionalities to be provided at each level of spatial resolution.
- Freedom of Information Act (FOIA) issues are of great concern to some data providers and many federal partners, and participants validated a web services approach to data dissemination that has the potential to provide data access without the implications of physical possession.
- Due to the wide range in data provider access policies, the workshop validated the approach of having local control over user authentication and access rights. Participants were impressed with the rights management processes employed by the British National Biodiversity Network (NBN), which provides administration tools, allowing local data providers to establish and maintain users' access to their data.

Complete proceedings of the workshop can be found online:

<http://www.cnhp.colostate.edu/projects/nsf/workshop-2004/index.html>.

### *Review of Comparable Efforts*

A review of similar online data access efforts has been undertaken to identify existing technological and sociological approaches to the issues facing the project. We are specifically interested in whether programming or other approaches can be adopted or repurposed for use in this project. Of particular interest are online data access systems currently in place across the NatureServe Network. Colorado State University has catalogued these efforts, and evaluated them with respect to the specific issues facing the NatureServe network. Special attention has been paid to the work of the British National Biodiversity Network (NBN), which has developed access control management approaches for use by a similar observation and monitoring network. Based on requirements gathered from the data access workshop, our software team developed prototype screens using an approach very similar to the NBN access control management system. These prototype screens were reviewed with the network of data providers for

initial reaction, and a working prototype of the security administration tools will be developed during year 2 based on this feedback.

#### *Research into Access Control Mechanisms*

The UMASS-Boston team has been researching existing XML-based access control standards and practices and evaluating their suitability for meeting the specific requirements of this project. The selected security architecture must be capable of serving either a central or distributed architecture. Among the access control approaches that have received particular scrutiny is a U.S. EPA-developed system that forms the basis for trading partner interactions over the secure Environmental Information Exchange Network.

#### *Enterprise Geodatabase*

Establishing an enterprise geodatabase to hold the aggregated polygon data describing species localities and extents was a prerequisite to creating a web service approach for serving this geospatial data online. Our enterprise geodatabase became operational in December 2004, using ESRI's SDE technology deployed on an Oracle platform. Procedures have now been established and utilities written to ensure that the geodatabase is synchronized with tabular data during the regular data exchanges with local nodes. The enterprise geodatabase is now serving as the primary spatial repository for the more than 500,000 element occurrence records managed by NatureServe.

#### *DiGIR Portal*

While the goal of this project is to establish full online access NatureServe Network data, as an interim step we established a DiGIR-compliant database to make a thin view of the data available through the Global Biodiversity Information Facility's (GBIF) prototype data portal. This required undertaking a cross-walking of Darwin Core data fields with data fields contained in NatureServe's Biotics 4 software system, and establishing a database view of the relevant data fields from which the DiGIR software can draw. We have had a number of conversations with GBIF and others involved in DiGIR and Darwin Core implementation about limitations of the current DwC version in supporting the type of geospatial and observational data managed by NatureServe (e.g., issues related to the "bounding box" and "basis of record" DwC fields). NatureServe's DiGIR service was made publicly available in February 2005. This initial web service can be discovered through the GBIF prototype data portal (<http://www.gbif.net/portal/index.jsp>); the Canadian Biodiversity Information Facility (CBIF) portal (<http://www.cbif.gc.ca/>); and the U.S. National Biological Infrastructure (NBII) portal (<http://gbif.nbio.gov/search/search.html>).

#### *Use Case Formulation*

Based on existing data requests to NatureServe and its state natural heritage program partners and on projected user needs gathered from interviews with NatureServe staff, member programs, and selected data users, a research team has been developing use case scenarios for developing detailed system requirements. Because many of the sociological issues involved in data sharing pertain to concerns that data will be inappropriately used,

this team has also begun developing *mis-use* cases as a means to drive discussions with data providers about what they are most concerned about, so that technological solutions can be developed that help them overcome those concerns and ultimately allow freer access to data resources.

### *Web Services Framework*

NatureServe has established a web services architecture in our central office, comprised of the following components. The foundation of this architecture is a Java-based web services framework that presents SOAP 1.1 and REST-based services, deployed using Apache Axis. The core data model for this framework is implemented using an Oracle and ESRI SDE enterprise geodatabase, including spatial representations of the element occurrence data aggregated from all network data providers, along with selected tabular attributes as needed to meet requirements of our initial user stories; an We have implemented one prototype web service using this framework to date. The target audience for this initial service is the network of data providers, so it will include a basic security interface to prevent unauthorized access. The service gives data providers access to range-wide element occurrence information for a given taxa, in order to promote access among the network of data providers to each other's occurrence data and to achieve benchmark data standards for element occurrence ranking information, which requires this range-wide view of how ranks are applied throughout the network.

## **YEAR 2 ACTIVITIES AND DELIVERABLES**

### *Data Exchange Workflow*

Ensuring taxonomic reconciliation and geospatial locational data synchronicity among the central and distributed nodes in the NatureServe Network currently is a time-limiting factor for accomplishing project goals. The data exchange team is researching processes for automating this workflow, and has adopted a goal of achieving 95% overall synchronization between the central publishing geodatabase and distributed local databases. This team has been evaluating the existing data exchange business processes and evaluating how to shift from a time-based exchange cycle to a data divergence-based updating system. The team is currently working on defining baseline measurement of percent synchronization, and beginning to refine and prioritize database fields that might trigger an automated exchange. The team is currently documenting high-level workflows for both taxonomic reconciliation and field-level data synchronization. Next, the team will select pilot states and data sets to begin experimenting with these new data reconciliation approaches.

### *Access Control*

The Access Control group at UMASS-Boston has concluded that a flexible Role Based Access Control architecture is most suited to the diverse disclosure limitation requirements of the natural heritage partners. It is therefore important to formulate precise requirements to include such concepts as under what circumstances data sharing partners would delegate role authentication to other partners or to a central authority. It is necessary to examine what are the architectural implications for machine access to disclosure limited data (e.g. support for local caching by web services clients), and what

tradeoffs in administration are necessary to support the wide diversity of technical expertise among the data providers. The UMASS team is working together with the NatureServe technical team and the Institutional Relationships team to develop a strategy for finalizing the detailed requirements to implement the access control module during year 2. A series of demonstrations accompanied with survey feedback forms will be used to develop a working prototype system and eventually the production-ready system.

#### *Develop a Menu of Web Services for Biodiversity Data*

NatureServe is developing a menu of standard and GIS-based web services that will respond to application requests. Examples of the type of request-response services under development are listed below. For responses that include location information, detailed or generalized map features will be returned in a format that can be displayed and integrated into open GIS applications.

- Submit the name of a species and retrieve detailed information about that species, including its legal and conservation status.
- Submit the name of a species and retrieve all known population occurrence locations for that species in North America.
- Submit a boundary for your area of interest, and retrieve a yes/no response about whether there are any known occurrences for threatened and endangered species in that area.
- Submit a species name and a boundary for your area of interest, and retrieve all known population occurrence locations for that species within the provided area.
- Submit a boundary for your area of interest, along with an Endangered Species Act status or NatureServe conservation status, and retrieve a list of the species matching your criteria known to occur in the provided area.

#### *Enhance User Interface to Geospatial Data*

The user interface team is working closely with the web services team to develop requirements for an improved user interface to the spatial data now available in the enterprise geodatabase. By the end of the second project year, we hope to have rolled out an improved NatureServe Explorer interface that supports spatial searches, data downloads and presentation of the enterprise geodatabase content. As web services are developed and published, example application interfaces will also be made available, in order to help jump-start clients with development of their own custom interfaces.