

Developing a New Infrastructure for Dynamic Access to Multi-institutional Biodiversity Data

PROJECT SUMMARY

This project will improve online access to the biodiversity data resources held by NatureServe and its network of natural heritage data centers, which collectively are the nation's leading source for detailed data on rare and endangered species and threatened ecosystems. Electronically confederating this distributed network of state biodiversity databases will promote open access to these data sets for the research and education communities and help inform conservation and environmental management. The project builds on work already carried out in developing and implementing a robust data model, creating technological tools for data management and data publishing, and addressing key institutional and intellectual property rights issues. The technology framework uses XML Web Services as a programmatic interface among four distinct technology layers: a gateway site, an enterprise geodatabase, an authentication/access control subsystem, and the distributed local databases. 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.

Intellectual Merit

This project will create enabling technology to promote open access of data from the NatureServe network's 75 participating institutions. By making large amounts of previously inaccessible data on the status and distribution of species populations and ecological community stands available to the research community, the project will create new opportunities for data exploration, analysis, and synthesis. New applications of these data will help advance scientific understanding of fundamental patterns and processes of the nation's biodiversity. This new infrastructure is also designed to promote interoperability between NatureServe's distributed database network and other major informatics initiatives, such as the Global Biodiversity Information Facility (GBIF). At present, the spatial representation of biodiversity in most distributed database initiatives is fairly basic. The rich geospatial attributes inherent in the NatureServe network's data provides an opportunity for this project to help advance general efforts to integrate standards and best practices from the geographic information systems (GIS) community with those of the taxonomy/collections community.

Broader Impacts

Improving online access to these unique biological data resources will have significant broader impacts by: 1) promoting conservation and sustainable environmental management; 2) advancing national goals for electronic government (E-Gov); and 3) enriching educational opportunities at K-12 and college levels. Natural heritage data are widely used for conservation planning and for environmental regulation and management. The improved access to these data provided by this project will substantially enhance their application in meeting societal goals for biodiversity protection and sustainable development. Most natural heritage programs are operated by state government agencies, and enabling online access to their data will improve government service to businesses, citizens, and other agencies, representing a major contribution to the nation's E-Gov infrastructure. The educational community at both K-12 and collegiate levels is already a major user of NatureServe's web resources, and the enhanced data and functionality available through this project will open up new opportunities for biodiversity information to serve as the basis for innovative lesson plans and geographically oriented student research. The project also seeks to build capacity in underrepresented communities by providing training opportunities at institutions with significant minority enrollment.

I. INTRODUCTION

The deteriorating condition of the nation's biological diversity is widely recognized as a major societal issue (Raven and Williams 2000). Stabilizing and reversing these declines while contributing to the sustainable use of biological resources will necessarily involve many sectors of society. The scientific community has a central role to play in documenting and understanding the species and ecosystems that represent the basic units of diversity, and for curating and disseminating the resulting knowledge to achieve broad societal benefit.

NatureServe is a non-profit research organization dedicated to providing scientific and technological support for conservation and environmental management. At NatureServe's core is a distributed database network consisting of natural heritage programs and conservation data centers that operate in all 50 U.S. states, the Navajo Nation, Puerto Rico, U.S. Virgin Islands, all Canadian provinces, and eight Latin American countries. Most U.S. participants in this public-private partnership are operated by state government agencies with natural resource responsibilities. NatureServe and its member programs focus on assessing the conservation status of species and ecosystems, documenting their localities through targeted field inventories, and managing spatially explicit databases that emphasize those species and ecosystems at greatest risk (Stein and Davis 2000). These data are widely used for conservation, land management, and regulatory decisions (Groves et al. 1995).

The creation of networks for integrating and disseminating biological data is a major scientific priority (PCAST 1998, Edwards et al. 2000), and a number of initiatives are underway in both the United States and globally that are designed to establish interoperable networks of biodiversity information (e.g., GBIF, NABIN, NBII). Similarly, using the Internet to provide improved government service—known as electronic or digital government, or E-Gov—is a priority at federal, state, and local levels (NRC 2002).

In operation for nearly 30 years, the NatureServe network is widely regarded as one of the most well-established and fully operational examples of a multi-institutional, geographically distributed biodiversity information network. NatureServe has made virtually all its centrally developed and managed data publicly available online (www.natureserve.org/explorer). The most spatially explicit data—precise localities for rare and endangered species and unusual ecological communities—are managed primarily at the state-agency level, and with a few exceptions these data are not available online. This lack of online accessibility greatly limits the broader potential of this resource for meeting the needs of the research and education communities, and for informing conservation and environmental management.

II. PROJECT GOAL

This proposal seeks to improve access to the proven biodiversity data resources held by NatureServe's network in order to increase their utility to researchers, educators, and environmental managers. We propose to accomplish this by developing and implementing a new infrastructure for data access that will electronically confederate this existing community of biodiversity databases. Electronically linking these largely state-agency databases through a Web Services architecture will provide for enhanced online data exploration, extraction, and analysis, and constitute an important contribution to the nation's digital government infrastructure. This project will build on work already carried out by NatureServe and its partners on developing and implementing a robust data model, creating technological tools for data management and data publishing, and addressing key institutional and intellectual property rights issues.

This proposal is a re-submittal of NSF proposal #0237697 submitted in July 2002, and addresses the issues raised by reviewers of that proposal (see Section IX for summary of comments).

III. BACKGROUND

The creation and maintenance of biological data networks involves an interplay among data, technology, and institutional issues. In 2000 NatureServe and Colorado State University held an NSF-funded planning workshop to explore the priority issues and constraints involved in establishing a distributed data access system for the network of natural heritage programs, and to design a conceptual framework for such a system. Attended by 79 technologists, scientists, and administrators from 45 private organizations, government agencies, and universities, the workshop participants were enthusiastic about the prospect of creating a system for improving online access to natural heritage data, and agreed that such a system would greatly benefit many users.¹ The workshop participants highlighted the importance of creating technological solutions to promote participation through incorporating appropriate data access and security measures, and the importance of developing institutional agreements designed to establish trust and ensure mutually beneficial relationships.

Following up on the results of this workshop NatureServe and its partners have invested considerable effort in clearly defining roles and responsibilities of network participants to ensure effective collaboration and data sharing. These roles are embodied in a new Data Sharing Agreement (DSA) negotiated between NatureServe and its member programs that builds on lessons learned over the past five years under an earlier agreement. Consistent with norms established by the Global Biodiversity Information Facility (GBIF), this new agreement encourages open access to the program's spatially explicit data, but affirms the rights of member programs as data custodians. The technology infrastructure proposed here is a critical enabler for meeting the goals of this new data sharing and access agreement, and for promoting open access to these data sets for researchers, educational users, and environmental managers.

Data, Data Models, and Metadata

NatureServe's primary goal is to improve the availability and use of biological and ecological information for informing conservation and land use decisions. Fundamentally, information is required that answers three questions: *what exists, how is it faring, and where is it found*. The information required to address these questions is encapsulated in a data model that has been the focus of refinement and evolution over more than 25 years, and is supported by a set of inventory and data management standards and protocols adhered to by network participants.² Element-referenced objects incorporated in the data model include information that relates to a species' (or community's) identity (including name and classification), status, general distribution, and life history characteristics. Spatial entities in the data model include the location and bounds of a species population or community stand, sites of ecological, scientific, or conservation interest, and areas under protective management.

Of particular significance is the concept of the *element occurrence (EO)*, the spatial representation of a species or ecological community at a specific location, and the primary unit of record in NatureServe's data model. An EO generally delineates a species population or ecological community stand, and represents the geospatial feature of biological interest. EOs are not synonymous with collection records. Rather, these records are spatial units documented in both space and time by voucher specimens and other forms of ecological observations. For example, a single plant population EO may be documented by multiple voucher specimens, each taken from different parts of the population, or from the same place over multiple years. While mapped objects can be polygons, linear features, or points, all EOs are represented as polygons to incorporate estimates of locational uncertainty. More than 500,000 spatially explicit EO records are managed across the NatureServe network representing several million observations or specimens.

¹ A summary of this workshop is available at <http://www.cnhp.colostate.edu/Projects/NSF/docs/workshop.html>.

² The full data model is publicly available at http://whiteoak.natureserve.org/hdms/Biotics_4-DataModel.shtml.

Because of the widespread use of EO data in land and natural resource management and regulatory decisions, inaccuracies can be costly. NatureServe's data development and management approach therefore is designed to minimize Type 1 errors. A detailed and rigorous set of scientific methods has been developed for documenting and mapping element occurrences, which incorporates estimates of uncertainty and accuracy and includes three distinct quality assurance steps.³ Metadata for each record allows users to select records that meet specific requirements for precision, currency, or type of documentation.⁴

Use of common inventory, data management, and taxonomic standards across NatureServe's network ensures that data gathered and managed by one network node is thematically and semantically comparable and electronically compatible with data gathered elsewhere in the network. This internal consistency enables the exchange and aggregation of data from multiple institutions across political boundaries, allowing regional and national-scale analyses and applications (e.g., Stein et al. 2000).

Technology and Software Tools

NatureServe's technology strategy focuses on developing tools that enable network participants to capture, manage, and disseminate biodiversity data, and to transform these data into conservation-relevant information. Encapsulating network-wide standards for inventory, mapping, and data management within these software products has been key to maintaining a coherent multi-institutional data set.

Biotics 4 represents the eighth generation of data management software developed for use by the NatureServe network, and was released in November 2002. Implemented in an Oracle database, the system integrates applications for spatial data management, tabular data management, data import/export and reconciliation, and reporting.⁵ The spatial component of the system is a custom geographic information system (GIS) application built on ESRI software, and supports digital mapping, spatial analysis, and data visualization. *Biotics 4* has been designed to provide a strong foundation for the Web Services architecture proposed here. Its Oracle platform supports Web service standards, and the data management interface, import/export, and reconciliation tools are all built on an XML-based data format. Installation of *Biotics 4* and conversion of legacy data to its new geospatial standards is an important step in web-enabling local databases. As of July 2003, 18 programs had installed *Biotics 4* and converted their legacy data, and most U.S. programs are scheduled for conversion by the end of 2004.

NatureServe Explorer (www.natureserve.org/explorer) represents our second-generation application for Web-based data publishing and exploration. The tool enables users to query approximately 50,000 species (including infraspecific taxa) and ecological communities by any combination of scientific or vernacular name, taxonomic group, conservation or legal status, and geography. NatureServe's public web offerings receive more than 60,000 visits (and 3 million hits) per month. Drawing from an Oracle back-end database, the NatureServe Explorer interface application is based on open source technologies, including Java servlets, Apache web server, and WebMacro servlet framework. Planned enhancements include the incorporation of an Internet mapping functionality to better serve the spatial data already offered on the site, and to provide a delivery vehicle for the network-wide EO data that is the focus of the current proposal.

³ Full documentation for the EO Data Standard is available at <http://whiteoak.natureserve.org/eodraft/index.htm>

⁴ FGDC-compliant metadata is available at <http://whiteoak.natureserve.org/hdms/Biotics 4-DataExchange.shtml>.

⁵ An overview of *Biotics 4* is available at <http://www.natureserve.org/prodServices/biotics.jsp>.

Institutional Arrangements for Data Access

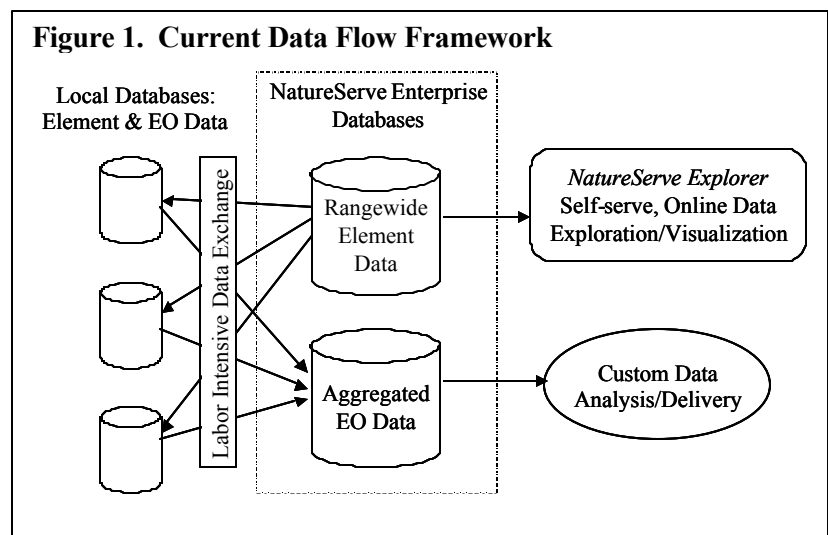
Data access policies vary across the network, with some programs making available complete GIS coverages of their most precise data and others bound by legal constraints on the provision of precise data. Many network programs consider at least some precise locational data to be sensitive, releasing them only on a need-to-know basis. The primary data sensitivity concern involves the risk that publishing precise locations of some rare species will expose them to poaching or deliberate destruction. A black market exists for many rare wild species, and nests of certain raptors, or populations of desirable orchids, cacti, or reptiles can be threatened by collectors. In other instances, private landowners may have motives to engage in legal or illegal activities designed to destroy endangered species populations or alter habitat. Additional concerns, especially in states with large rural areas, involve the rights of private property owners to privacy regarding the endangered resources on their land. State governments take the obligation to respect these rights very seriously.

As a membership organization, NatureServe's members have a governance stake in the organization, and committee and work group structures are in place to collectively address issues ranging from institutional priorities, to development of data standards and technology, to data access. Over the past several years NatureServe and its member programs have made considerable progress in creating the institutional framework for aggregating EO data into a national-level database and for making these data more broadly accessible. A new Data Sharing Agreement (DSA) was negotiated over the past year between NatureServe and each of the network participants, which has the goal of encouraging open access to fine-scale spatial data. Whereas the first generation of data sharing agreements negotiated five years ago strove to create a nationally consistent "least common denominator" approach to data sharing, the current agreement seeks to maximize the level of access provided by each program. For each data custodian the DSA specifies the terms under which access to such precise data may be provided. Important issues already addressed in the data sharing agreements relate to commercial versus non-commercial uses of data, appropriate acknowledgements and attributions, re-dissemination to third parties, and the appropriate uses of sensitive information (NRC 2000).

Letters of participation from more than 35 natural heritage data centers are attached, demonstrating the widespread enthusiasm among network participants for developing online data services designed to provide more open access to their data, and to meeting the needs of the broader research community.

Current Data Flow

Figure 1 diagrams the network's current data flow, where range-wide element data are made available for public dissemination via NatureServe Explorer, while more precisely georeferenced multi-state EO data are available only through custom data processing and delivery. The network realizes economies of scale by developing range-wide data centrally and distributing them to local nodes through regular data exchanges. While NatureServe collaborates with ITIS and others in maintaining standard taxonomies for use across the network, state programs are free to adopt



alternative taxonomies that meet local preferences or requirements (e.g., taxonomic concepts embodied in state regulations) (Morse 1993). The coupling of needed taxonomic reconciliation with routine data exchanges, however, limits the frequency of such exchanges to once a year per node.

Local programs are the primary data custodians for EO data and serve as the principal distribution agents for that data within their states. Although individual centers largely meet the demand for data within their own jurisdictions—answering more than 75,000 information requests annually—until recently there has been no mechanism for providing access to EO data across multiple jurisdictions deriving from multiple institutions. The realization of formal Data Sharing Agreements with all network participants has enabled the creation of a national EO data set, managed at NatureServe’s central office and available for meeting the needs of multi-state users. Access to this important resource, however, is currently available only through custom processing and delivery (Figure 1), a situation this proposal is designed to address.

Interoperability with Other Biodiversity Networks

NatureServe collaborates with a number of other international initiatives focused on improving dynamic access to biodiversity data. We have joined the Global Biodiversity Information Facility (GBIF) as an associate participant—the category available for non-governmental members—and have been designated a thematic node of the Convention on Biological Diversity’s Clearinghouse Mechanism (CHM). We are also participating with and serve on the steering boards for both the North American Biodiversity Information Network (NABIN) and the Inter-American Biodiversity Information Network (IABIN), and are an institutional member of the Taxonomic Database Working Group (TDWG). At the U.S. level, NatureServe is a partner of the National Biological Information Infrastructure (NBII), and the only non-governmental member of the Integrated Taxonomic Information System (ITIS) partnership.

A major goal of the current proposal is to improve the ability to make the NatureServe network’s data resources interoperable with these initiatives. Several standards and tools are emerging from the various initiatives that are relevant to the work proposed here. The ecological informatics community is involved in developing a domain-specific metadata schema, known as EML (ecological metadata language), which has relevance to types of information managed by NatureServe. The Access to Biological Collections Data (ABCD) task group is in the process of developing a collections-oriented database schema as well as a protocol (DiGIR) for retrieving structured data from multiple heterogeneous databases. Of particular significance is the recently created TDWG Spatial Data Standards subgroup, which is focusing on the problem of integrating existing standards and best practices developed by the geographic information systems (GIS) community with the practices and needs of the taxonomy/collections community, and evaluating uses of XML for access to distributed spatial databases. The intersection between geospatial and taxonomic/collections standards and practices lies at the heart of NatureServe’s efforts to provide access to geospatial biodiversity data, and we will be actively engaged in this subgroup to help advance that effort, and to accomplish our project objectives in a way that maximizes interoperability with other initiatives.

Training & Technical Support

A serious commitment to training is central to maintaining the network’s effectiveness. The core element of this training program has been a formal weeklong course focusing on inventory, mapping, and data management standards and protocols. Offered on a regular basis, the most recent course represented the 106th such training session. Formal training opportunities are also offered at regular regional conferences for network participants and at NatureServe’s annual meeting. Technical support for data management applications occurs at many levels. Assistance in deploying new software (e.g., Biotics 4) includes data conversion support, application installation and configuration, and on-site training. The software itself includes substantial on-line help, while NatureServe support services

include a Web-based knowledge base, telephone support, and email discussion lists. Application training is also conducted via the Internet utilizing technology that allows remote users to share a common desktop.

IV. PRIOR NSF SUPPORT

Award: NSF ITR/IM(BIO) #0113058 **Amount:** \$497,0370. **Period:** 9/01/01-08/31/04. **Title:** *Biodiversity Data Discovery and Integration*. **PI:** Robert Morris; **Co-PI:** Robert Stevenson. This project builds a component architecture by which integration of biodiversity web information can take place seamlessly. It introduces building blocks to locate the origin and authority of a species name, to extract maps from servers that can compute species distribution, and provides “nuts and bolts” by which unrelated databases and web sites can be made to cooperate. This is accomplished through use of XML, and Web Services, especially framework tools such as Apache Axis, Apache Web Services Invocation Framework, and Java Databinding tools such as Exolab’s Castor framework. We have built Web Services layers around ITIS, around our own electronic field guides, and are producing a wrapper around Robert Colwell’s BIOTA software that will permit BIOTA users to participate in GBIF and other architectures using the TDWG DiGIR protocols. We have also produced an invasive species ontology in DAML+OIL, the Darpa Agent Markup Language+the Ontology Interface Layer. Morris, R. A. and R.D. Stevenson. In prep. Integrating heterogeneous biodiversity applications.

Proceedings of the Workshop on Ecoinformatics, Bangalore, India, June 2003.

Morris, R. A., R. D. Stevenson and W. Haber. Submitted. Architecture of Electronic Field Guides. Stevenson, R. D., W. A. Haber, and R. A. Morris. 2003. Electronic field guides and user communities in the eco-informatics revolution. *Conservation Ecology* 7(1): 3. [online] URL: <http://www.consecol.org/vol7/iss1/art3>.

Morris, R. A., M. Passell, and R. D. Stevenson. 2001. A Software Engineering Perspective on Developing Electronic Field Guides: Lessons Learned For Bioinformatics. European Environmental Agency Technical Report Series.

V. PROJECT APPROACH

Building on progress already made in establishing a robust data model and addressing key data sharing issues (e.g., negotiation of a new generation of DSAs), our project approach focuses on creating the enabling technology that will promote open access of data from network participants. Our proposed technology framework uses XML Web Services as a programmatic interface among four layers: a gateway site, an enterprise geodatabase, an authentication/access control subsystem, and the distributed local databases.

- Gateway Site. The gateway site will provide the primary user interface, allowing data exploration, visualization (mapping), and extraction. This gateway will build on the existing NatureServe Explorer application and serve as a thematic index site, allowing advanced queries based on taxonomy, geography, and conservation/legal status, as well as other attributes important to users (e.g., habitat preferences).
- Enterprise Geodatabase. An enterprise geodatabase will provide the framework for managing the large volumes of network-wide geospatial EO data, and will support queries against the gateway. This database will enhance multi-jurisdictional data presentation and analysis and provide crosswalks between local and network-wide taxonomies.
- Authentication/Access Control Subsystem. This subsystem will enable us to apply rights management transformations to data served in XML by the enterprise geodatabase and distributed local databases. Separating the rights enforcement engine from the data will provide for greater scalability and separation of data retrieval and rights enforcement.

- Local Databases. More than 75 local databases support local-level data input, quality control, management, and dissemination, and are the primary repository for EO geospatial data. Web-enabling these nodes will ultimately allow for direct responses to queries received from the gateway site, as well as allow interoperability with other sources through support of third party applications.

Variation in technological and financial capacities among network programs requires a solution that is flexible at the local level, and does not immediately require web-enabled local databases. Our project approach (Figure 2) is phased in such a way as to first support queries through an enterprise geodatabase (Version 1), and later implement a fully distributed architecture for querying local databases (Version 2). The current project will focus on delivering Version 1, since available project funding does not allow for implementation of Version 2.

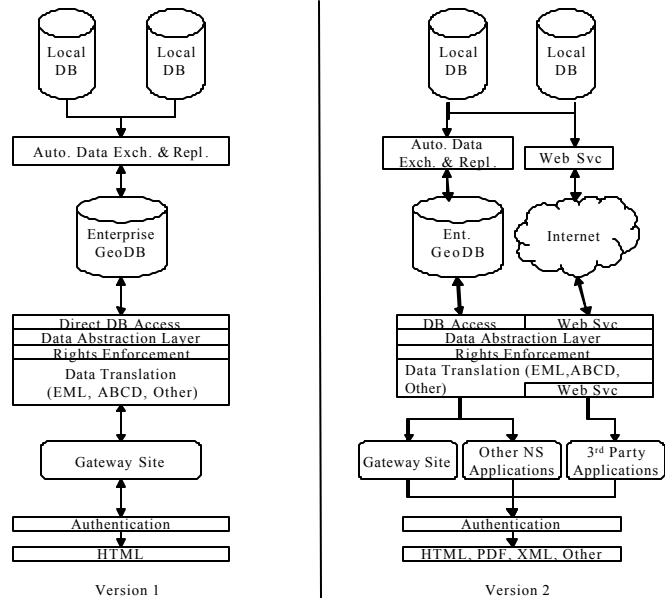
Version 1 will establish a gateway site expanding on the existing NatureServe Explorer application, create an enterprise geodatabase containing a replicated set of all network EO records, and develop data access/authentication components based on codification of the DSA. During the development of Version 1 we will increase the installation rate of Biotics 4 software in network data centers, since moving onto this Oracle-based system will greatly facilitate making their spatial data more efficiently available over the Web.

Future implementation of Version 2 would deploy XML Web Services technology to enable direct access to those local databases that are Web-enabled. Data from network participants not yet Web-enabled will continue to be accessible through the enterprise geodatabase. This hybrid solution is designed to ensure that where possible users access up-to-date information directly from the data custodian, yet are still able to access data from across the entire network. The enterprise geodatabase therefore plays an important role in providing short-term access to network-wide data and longer-term redundancy to ensure robust operation and adequate system performance. Over time we foresee transforming the enterprise geodatabase from a replicate EO database containing network-wide spatial data to a collective mirror site that receives refreshed local data on a continuous basis, and which provides a back-up when one or more distributed sites are unavailable.

We will leverage technology in several areas: 1) distributed database query and access using a loosely coupled, n-tier architecture; 2) Internet mapping services and advances in data streaming that allow the dynamic transfer and assembly of very large geospatial data files; and 3) established XML Web Services standards that provide protocols for description (WSDL, Weerawarana et al. 2002), discovery (UDDI, Bellwood et al. 2002), and communication (SOAP, Box et al. 2000). Wherever possible we will integrate open source or commercial off-the-shelf (COTS) software to minimize development costs and maximize availability of training and support across our institutionally heterogeneous and geographically dispersed network.

Our proposed technology framework (Figure 2) will create layers responsible for distinct tasks as data moves through the system to final delivery. Under this project (Version 1) we will focus on

Figure 2. Concept Diagram for the Project Approach



creating the enterprise geodatabase and the data access/authentication components. These components will include authentication, data abstraction, rights enforcement, and data translation. Data will move between each component layer in XML documents. In addition, we will create tools for automating the data exchange process between local databases and the enterprise geodatabase by separating it from the taxonomic reconciliation process, thus allowing for more frequent exchanges. A future enhancement (Version 2) will involve delivering XML Web Services to enable a network node to provide direct access to its data from the Gateway site or from other compatible third party applications. This phasing will allow us to create the critical pieces of data aggregation, translation and access control as part of this project funding, even as we seek other funding to subsequently deploy these via Web Services.

The project also requires seamless access to the geodatabases utilizing Web Services and access control technologies. This is a period of rapid change in the geospatial technology community, and we are closely following efforts to utilize open standards for Internet-based geoprocessing, and the technology developments from both the Open GIS Consortium and ESRI. We will be looking particularly closely at ArcGIS 9.0, which is expected to provide loss-less XML serialization of spatial data, replication and a standard XML schema for GIS data (GML).

Web Services for Providing Interoperability

XML-based Web Services will be used to make local data accessible to the broader community through the gateway site, using emerging standards to maximize interoperability with other applications. The services will be a part of the Access/Authorization Components. These components include a data abstraction layer that will support spatial and tabular queries. Data abstraction will encapsulate the enterprise geodatabase data model and make it available in a standard XML schema through an object model. The abstraction layer will use Web Services deployed in network locations to retrieve data from multiple sources. Full implementation of this distributed database approach will require that each of the local databases be Web-enabled and capable of responding to gateway queries. The Web Services will be based on the Apache open source tools (<http://ws.apache.org/>), principally the Axis Web Services framework. The result will serve both XML, valid for an XML-Schema designed in collaboration between the University of Massachusetts-Boston team (UMASS) and NatureServe, and HTML consistent with current interfaces but produced from the XML by the application of XSLT the eXtensible Style Sheet Transformation language. This allows HTML to be produced against XML that has gone through the rights enforcement layer. Client side support will be based on the Apache Web Services Invocation Framework.

Authentication and Access Control

The authentication and access control subsystem will provide a means by which the most sensitive data can be made available to permitted classes of users. Under the direction of Dr. Robert Morris, the UMASS team will develop the access control (AC) layer. The team will design and implement an XML access control system interposed between the backend and the Web Services layer. It will be a role-based system (<http://csrc.nist.gov/rbac/>). Hence, we propose a fine-grained XML AC system based on roles in an optional signed digital certificate accompanying queries. A data provider offering XML will place into all sensitive XML elements, roles and policies for those roles expressed in the XML Access Control Language (XACL)(www.trl.ibm.co.jp/projects/xml/xacl/index.htm). This enhanced XML is passed to an Access Control Agent (ACA) which may be a separate web service or may be directly invoked by the provider as we propose in Version 1 implementation. Rights enforcement will apply security transformation to the XML data documents based on stored roles, as codified in the DSA. If the roles certified by the certificate accompanying the query match those in the sensitive element, the policy (including policy for the absence of a certificate) will be enforced by the ACA, which returns the resulting XML to the requesting client, perhaps with further processing

(e.g. conversion to HTML if the client is a browser-based application). Examples of some roles and policies that might be supported: SeniorScientist: pass precise geospatial data; RegionalPlanner: convert precise geospatial data to appropriate place name. We note that the architecture is fully dynamic, permitting the data provider to change the policy over time, consistent with obligations made in their DSA. The implementation will provide middleware for inserting XACL at the source, as well as for implementing policies in the ACA. We will adopt the design expressed in the Trust Establishment (TE) architecture described by IBM (<http://www.alphaworks.ibm.com/tech/TrustEstablishment>). This system permits digital certificate issuers to authenticate a holder to a role, and obviates the need for password maintenance or individual identity authentication.

To be successful this approach must provide easy and transparent data access from the user perspective, but must also respect and support the data access policies of the individual data providers. Technologically, the system will be designed to provide the finest level of access that an individual data provider allows, employing user-role authentication XML access control. At the same time, we believe this project will create the institutional buy-in and peer pressure required to increase the amount of fine-scale geospatial data available for scientific, educational, and other use by working with network participants and others to identify and address key institutional and intellectual property rights issues. A workshop in year one will focus on codifying user role and policy descriptions to be embedded in the authorization and access control subsystem.

VI. PROJECT TEAM AND COLLABORATIONS

Two teams, working under the direction of the principal investigator, will implement the project by focusing on specific themes and deliverables. We are also committed to tapping into innovative technology work taking place across the network of natural heritage programs, and have established a network-wide Technology Working Group (TWG) to help empower local technologists and harvest these innovations for broader application. Because a key objective of this project is to promote interoperability with other key initiatives, we have established a Project Advisory Committee that reflects multiple external perspectives, and will provide guidance to the PI and Co-PIs on project direction and implementation.

Advisory Committee (AC)

Lead Investigator: Bruce Stein

A Project Advisory Committee has been assembled to provide external perspectives from individuals involved in various aspects of biodiversity and ecological informatics. The following individuals have agreed to serve on this Advisory Committee (see attached letters): Dr. James Beach (University of Kansas); Dr. James Brunt (Long Term Ecological Research Network, University of New Mexico); Dr. Frank Davis (University of California, Santa Barbara); Dr. Steven Kelling (Cornell Laboratory of Ornithology); Dr. Meredith Lane (GBIF Secretariat); Dr. Ronald Pulliam (University of Georgia); and Dr. Barbara Stein (MaNIS project, University of California, Berkeley).

Institutional Relationships Team (IR)

Lead Investigator: Mary Klein

Purpose: Work across institutions to involve data custodians in the design of system components and control processes that embody agreed-upon data sharing and access guidelines. Identify workshop participants and host data access workshop. Document participant and user requirements for authentication and access control system.

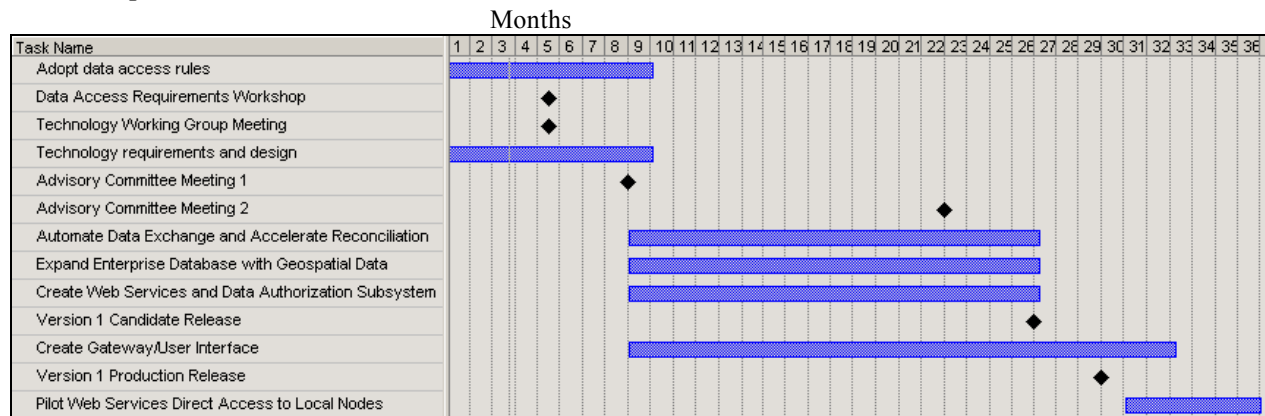
Technology Design and Implementation Team (TDI)

Lead Investigators: Larry Sugarbaker and Robert Morris

Purpose: Design and implement the technical architecture and software components.

- Gateway Site Team – design user interface and program backend changes to support spatial searches from the geodatabase.
- Web Services Team – design and develop the Web Services and access control components.
- Data Model Team – design the enterprise geodatabase.
- Technology Working Group – consists of a team of technologists from data centers across the NatureServe network that have expertise or interest in online data access technologies. Will serve as a network-wide collaboratory fostering innovations and providing input into design and implementation.

VII. Workplan



Adopt Data Access Rules

Existing DSAs provide detailed information about custodian data release policies, especially regarding data access levels and user types. This project will begin by clarifying the design requirements to embody these policies in automated technologies and ensure that requirements are understood for all cases. As needed, we will also update the DSAs to reflect the emerging needs of an Internet data delivery environment. Colorado State University’s Colorado Natural Heritage Program will engage students in hosting a workshop to review data provider expectations with respect to the automation of existing user-authentication agreements, and to document data user expectations regarding appropriate uses and ease of access. The workshop will include network participants as well as partners from the broader scientific community. The goal of the workshop will be to refine the requirements and online implementation of the data sharing policies embodied in the DSAs. Guidelines will document specific classes of users that can be used to form the basis for the data access/authorization subsystem.

Tasks	Deliverables
<ul style="list-style-type: none"> • Workshop to clarify and refine data sharing and delivery rules 	<ul style="list-style-type: none"> • Data access guidelines and user authentication rules published on the Internet • Design requirements • Refined data sharing agreements

Technology Requirements and Design

We will review infrastructure options and design the most appropriate architecture for meeting the needs of the user community. An appropriate XML schema will need to be derived for EO data. Technical design and interface specs will be worked out, with input from the CSU workshop guiding the final stages of development. We will also evaluate the feasibility of connectivity solutions for the Geodatabase based on an ESRI ArcGIS 9.0 platform.

Tasks	Deliverables
<ul style="list-style-type: none"> • TDI teams review of existing data delivery approaches, alternative data exchange work flows, and requirements • Evaluate suitability of potential software tools and recommend specific infrastructure requirements 	<ul style="list-style-type: none"> • System Design documents for all aspects of technology <ul style="list-style-type: none"> ○ Data abstraction layer ○ Access control agent specifications ○ Translation libraries ○ Modifications to Gateway interface and Biotics 4 ○ XML schema ○ Incremental data exchange protocols

Automate Data Exchange and Increase Frequency of Data Reconciliation

Regular data exchanges between NatureServe and network programs are required to ensure up-to-date range-wide and taxonomic data at the local nodes, and to aggregate local data into the enterprise geodatabase. These exchanges are complicated by the need to resolve taxonomic differences between local and central databases. Data exchanges currently are linked directly to the taxonomic reconciliation process, a labor-intensive scientific review process that limits the frequency of updates to once a year per local node. Automating the data exchange process and separating it from the taxonomic reconciliation process will allow information to be exchanged on a more frequent basis, greatly improving the amount of up-to-date range-wide data available to the local nodes. One possible approach is to create an incremental update process to enable more frequent updates. Incremental updates would only include data that have changed from a previous update. By reducing the amount of data to only the changes should allow for more frequent refreshes. We will also be reviewing taxonomic reconciliation software currently under development elsewhere to determine its applicability to our needs.

Tasks	Deliverables
<ul style="list-style-type: none"> • Build on Biotics 4 data exchange tools to automate synchronization of local and range-wide data, and identify records requiring review and reconciliation • Design a new work flow system for taxonomic reconciliation • Train network participants in new work flow procedures 	<ul style="list-style-type: none"> • Automated, incremental data exchange and synchronization tools • Revised taxonomic reconciliation process, including workflow and labor impacts • Methodology training for network programs

Expand Enterprise Database to Include Network-Wide Geospatial Data

The new enterprise geodatabase will integrate EO layers from all network nodes into a common database environment that will support data exploration, visualization, and extraction through the proposed gateway site. This geodatabase will manage the spatial representation of all EO data as polygons rather than as geo-referenced points. With Biotics 4 installed, each local database will have a single ArcView GIS layer containing their entire set of EO records, represented as polygons. These ArcView shape files will be converted to a more robust geodatabase format based on the guidance developed in the Technology Requirements and Design activity, providing the capability to effectively manage the large volume of network-wide geospatial data. Implementation will include tools and processes to integrate the exchange of geospatial data into the regular data exchange and reconciliation schedule.

Tasks	Deliverables
<ul style="list-style-type: none"> • Build enterprise geodatabase infrastructure including data dictionary, data model and 	<ul style="list-style-type: none"> • Enterprise geodatabase integrating network-wide spatial element occurrence data and range-wide

Tasks	Deliverables
file management protocols <ul style="list-style-type: none"> • Develop tools to support refresh of the enterprise geodatabase from local nodes • Implement a regular refresh schedule for geospatial data 	element data <ul style="list-style-type: none"> • Test plan

Create Web Services and Data Access/Authorization Subsystem

Data access control policies and requirements will be incorporated into the system using an Access Control Agent (ACA) programmed in the XACL language to support XML-based management of the rights of specified user classes. We will then develop Web Services components in Web Services Description Language (WSDL) using the Apache Axis framework and Web-Service Invocation Framework, in conjunction with server-side programs that will respond to queries from GBIF/TDWG DiGIR compliant users. Students from the University of Massachusetts-Boston will develop the ACA and WSDL code, playing a lead role in deployment, testing and documentation. The Advisory Committee will review proposed solutions for transparency, ease of use, and suitability of pre-defined roles for the research community.

Tasks	Deliverables
<ul style="list-style-type: none"> • Develop server-side facility for XML access control • Implement role certificate management subsystem with Public Key Infrastructure • Build Access Control Agent to transform controlled elements for delivery to users • Develop Web Services 	<ul style="list-style-type: none"> • User authentication and Access Control Agent • Web Services software components <ul style="list-style-type: none"> ○ WSDL compliant services using Apache Axis ○ Client-side invocation of Web Services also using Apache ○ GBIF/TDWG DiGIR compliant protocols ○ Metadata in EML registered where suitable (e.g., Metacat) • Test plan • Documentation for installing and maintaining the system

Create Gateway/User Interface

Version 1 will be delivered via an enhanced NatureServe Explorer user interface to support spatial searches, data downloads and presentation of the enterprise geodatabase content. Central node components from all preceding project activities will be tested and deployed via this interface.

Tasks	Deliverables
<ul style="list-style-type: none"> • Modify NatureServe Explorer interface to create a gateway to the enterprise geodatabase • Develop user-oriented documentation to describe appropriate uses and guidelines for interpretation of the data sets 	<ul style="list-style-type: none"> • Gateway site providing self-service, online data discovery, visualization, and XML-based delivery • Test plan

Pilot Web Services Direct Access to Local Nodes

Implementation of Web Services for direct access to local nodes (Version 2) is not a part of the current project funding. However, to ensure that Version 1 of the software architecture will meet that ultimate goal, we will work with two local pilot sites during the current project. Pilot sites will be selected based on their ability to provide training opportunities for underrepresented populations, or to help build institutional capacity.

Tasks	Deliverables
-------	--------------

Tasks	Deliverables
<ul style="list-style-type: none"> • Engage two network nodes to serve as pilot test sites • Test and deploy local node Web Services components in collaboration with pilot test sites • Train network nodes on the enterprise geodatabase data models and Web Services tools 	<ul style="list-style-type: none"> • Test plan • Two operational local Web Services nodes • Rollout plan for network-wide implementation over longer time frame • Technical architecture, system implementation and maintenance documentation • Long-term, network-wide training plan

VIII. BROADER IMPACT

Creating online access to this unique biological data resources will have broader impacts in promoting conservation and sustainable environmental management, in advancing electronic government (E-Gov), and in enriching education at K-12 and college levels. Additionally, the project will seek to involve graduate students from underrepresented communities in local pilots and build capacity in historically underserved jurisdictions.

Conservation and Environmental Management

Biodiversity conservation is a major societal concern, and sound scientific information is needed to set targets for protection and improve management of natural resources. Data from NatureServe’s network of natural heritage programs is widely used for conservation and environmental management; indeed, *The New York Times* has referred to the network’s databases as “the country’s leading source of biological information for conservation planners, government agencies and land managers” (Stevens 2000). While these programs collectively fill more than 75,000 data requests annually—informing decisions ranging from nationwide oil-spill contingency planning to the evaluation of site-specific impacts from housing developments—most of these requests are serviced manually. As a result, many users and potential users are not incorporating these data into their routine decision-making processes to the degree that would be desirable. Creating online access to these data should greatly expand the number of prospective users for this information, producing correspondingly broad environmental benefit.

Electronic Government

Electronic government, or E-Gov, is a national priority designed to provide improved government service to citizens and businesses by, among other things, sharing and integrating federal, state, and local data. Because most natural heritage programs are operated by state government, enabling dynamic online access to this information represents a major contribution to meeting E-Gov goals. Implementation of this distributed database access system will improve the efficiency and effectiveness of these state government programs by allowing them to spend less time on routine data requests and environmental reviews and focus more attention on consultations requiring in-depth biological expertise. Many users of the data are involved in environmental or development projects where delays in planning or permitting can have significant financial costs. By reducing the response time for data requests, this project should also represent a cost savings to the government agencies, businesses, and citizens that rely on these data. Developing an enterprise solution to improving online access to state natural heritage data—the focus of this proposal—will also create efficiencies and cost savings by avoiding the need for each state to create its own technological solution. Finally, such a network-wide solution will allow data to be queried and integrated across state boundaries, better meeting the E-Gov needs of regional bodies and federal agencies.

Education

The biodiversity data held by NatureServe and its member programs also have broad application in the educational community, including K-12. The NatureServe Explorer web site, which will be expanded to serve as the gateway for the new distributed architecture, already has been identified as a leading web-based educational resource. The American Library Association's review of NatureServe Explorer describes it as "a tremendous new resource that deserves to be bookmarked at every library, whether public, school, or academic" (ALA 2001). Similarly, NatureServe Explorer is highlighted as a "great site for middle and high school research" by Education World (www.education-world.com/a_sites/sites080.shtml), and is included in numerous local and national "homework help" sites, such as that maintained by the Carnegie Library of Pittsburgh (www.carnegielibrary.org/subject/homework/biology.html). This project is intended to greatly enhance the level of geographic information available to students through NatureServe Explorer, opening up new opportunities for this information to serve as the basis for innovative lesson plans and geographically oriented student research. In particular, this new data and functionality will open up opportunities for students to both explore their local environs and to compare and contrast this with ecosystems elsewhere.

Underrepresented Groups

The project will also strive to benefit underrepresented populations through their involvement in training and institutional capacity building. In particular, we intend to pilot local implementation of the Web Services infrastructure at university-based natural heritage programs with significant minority enrollments, such as University of New Mexico (44% minority) or University of Alaska (28%).

IX. COMMENTS FROM PREVIOUS SUBMISSION

The panel reviewing the previous submittal of this proposal agreed that the natural heritage databases are an invaluable resource and that confederating them and making them more accessible is an extremely worthwhile and important goal. The panel raised several issues regarding: 1) the level of detail presented in the technology plan; 2) development of access and authentication systems; 3) resolution of institutional issues related to data access; and 4) level of external participation to promote compatibility with other initiatives.

Technology Plan

Considerable detail has been added to the technology plan to lay out the conceptual basis for our approach and outline the technology solutions contemplated. While some of the technologies to be employed are quite mature and their implementation will be straightforward (e.g., Web Services), other standards and technologies are evolving rapidly, especially the field of Internet-based geoprocessing. Thus, a number of specific details must be deferred until the technology requirements and design phase of the project to ensure that we take advantage of the most appropriate emerging standards and technologies.

Access Control and Authentication

A detailed plan for developing access control and authentication is presented in the project approach and work plan sections of the proposal. This work will be undertaken by the UMASS-Boston team lead by Co-PI Robert Morris.

Institutional Issues

The formalization of institutional agreements among network participants has been essential for meeting the objectives of this project. Work on developing consensus among institutional participants on data sharing issues over the past three years has yielded significant advances, especially with respect to the documentation of data custodian expectations regarding control and release of information to authorized users. The creation of NatureServe in 2000 as an independent membership organization representing the confederated databases, including governance mechanisms involving network members, has greatly facilitated inter-institutional data sharing arrangements. As discussed earlier, over the past year NatureServe has worked with its member programs to implement a next generation of formal data sharing agreements (DSAs) that are directly supportive of this proposal's approach to improving access to detailed geospatial element occurrence data. Letters from more than 35 natural heritage programs are attached, indicating their commitment to participate in this project.

External Participation

Dr. Robert Morris of the University of Massachusetts-Boston has joined the project team as a Co-PI, and we will benefit enormously from his understanding of and involvement with many of the other leading biodiversity informatics initiatives. A specific goal of this project is to improve interoperability of the NatureServe network with other initiatives, and Dr. Morris will be working with NatureServe technology staff to develop a technical architecture designed to maximize such interoperability. We have also continued to include in the project an external Advisory Committee, with representation from individuals involved in museum database confederation projects (MaNIS, University of Kansas), international initiatives (GBIF), ecological informatics networks (LTER), and citizen-science web projects (eBird). These individuals will bring a varied external perspective to the design and implementation of the system. Letters from these seven external advisors are attached indicating their commitment to serve on this committee.

SECTION D: References Cited

American Library Association (ALA). 2001. Review of NatureServe Web site. *Choice: Current Reviews for Academic Libraries* 38(9): 3889.

Bellwood, T., L. Clément, D. Ehnebuske, A. Hately, M. Hondo, Y. L. Husband, K. Januszewski, S. Lee, B. McKee, J. Munter and C. von Riegen. 2002. Universal Description, Discovery and Integration (UDDI) Version 3.0. <http://uddi.org/pubs/UDDI-V3.00-Open-Draft-20020703.htm>. Web page accessed July 17, 2002.

Box, D., D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H. F. Nielsen, S. Thatte and D. Winer. 2000. Simple Object Access Protocol (SOAP) 1.1. WC3 Note 08 May 2000. <http://www.w3.org/TR/SOAP/>. Web page accessed July 17, 2002.

Edwards, J. L., M. A. Lane, and E. S. Nielsen. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* 289: 2312-2314.

Groves, C. R., M. L. Klein, and T. F. Breden. 1995. Natural heritage programs: Public-private partnerships for biodiversity conservation. *Wildlife Society Bulletin* 23: 784-790.

Morse, L. 1993. Standard and alternative taxonomic data in the multi-institutional Natural Heritage Data Center Network. Pp. 69-79 in F.A. Bisby, G.F. Russell, and R.J. Pankhurst eds. *Designs for a Global Plant Species Information System*. Oxford: Oxford University Press.

National Research Council (NRC). 2000. *The Digital Dilemma: Intellectual Property in the Information Age*. Washington, DC: National Academy Press.

National Research Council (NRC). 2002. *Information Technology Research, Innovation, and E-Government*. Washington, DC: National Academy Press.

President's Committee of Advisors on Science and Technology (PCAST). 1998. *Teaming with Life: Investing in Science to Understand and Use America's Living Capital*. Washington, DC: White House Office of Science and Technology Policy.

Raven, P. H. and T. Williams. eds. 2000. *Nature and Human Society: The Quest for A Sustainable World*. Washington, DC: National Academy Press.

Stein, B. A., L. S. Kutner, and J. S. Adams. eds. 2000. *Precious Heritage: The Status of Biodiversity in the United States*. New York: Oxford University Press.

Stein, B. A. and F. W. Davis. 2000. Discovering life in America: Tools and techniques of biodiversity inventory. Pp. 19-53 in Stein, B. A., L. S. Kutner, and J. S. Adams eds., *Precious Heritage: The Status of Biodiversity in the United States*. New York: Oxford University Press.

Stevens, W. K. 2000. U.S. found to be a leader in its diversity of wildlife. *New York Times*, March 16, 2000, p. A18.

Weerawarana S., R. Chinnici, M. Gudgin, and J. Moreau. 2002. Web Services Description Language (WSDL) Version 1.2. WC3 Working Draft 9. <http://www.w3.org/TR/wsdl12/>. Web page accessed July 17, 2002.