Integrating the Distributed Data Resources of the Bird Monitoring Community Using Information Technology Strategies[1]

Steve Kelling[2] and Craig Stewart[3]

Summary

An increasing number of bird monitoring projects are assembling massive quantities of data into numerous decentralized and locally administered storage systems. These data sources have enormous significance to a wide range of disciplines, but knowing that they exist and gaining access to them is difficult if not impossible. Attempts are being made to organize these data sources into a unified data resource, with the biggest challenge being to develop an infrastructure that facilitates data integration without compromising local control, and maintaining, when appropriate, the privacy of the data. Broadband networks provide a means to make this data available, and advances in hardware and software applications provide tools to organize and supply access to these data sets. This paper attempts to identify the key issues in organizing the contents of the bird monitoring data sets, and to discuss the technologies and solutions to overcome them.

Identifying the Problem

A major theme at the Third International Partners in Flight Conference, *A Workshop on Bird Conservation Implementation and Integration*, held 20-24 March 2002, was how to provide ways that encourage partnerships among various levels of government and non-government organizations interested in bird conservation. Beginning with a keynote address by David Brackett (this volume), Director General of the Canadian Wildlife Service, the Internet and Information Technologies (IT) were identified as playing key roles in providing the tools for integrated bird conservation. Developing an IT strategy that supplies the necessary infrastructure to unify the thousands of bird monitoring projects into a useable organization for eco-regional planning, to make biological assessments for bird conservation, and to provide a means for cooperative research and education is now a goal that can be achieved.

But there are challenges. Bird monitoring data are widely distributed across an unorganized array of data structures, many of which are poorly maintained, and have limited access. First, as time progresses there is a risk of losing valuable, even unknown, data unless the bird monitoring community comes together with a common goal of preserving and making this data available. Second, while it is imperative to organize these resources into a federated system that provides access to the data, the system must not compromise local control, and must ensure the necessary privacy of sensitive data. Finally, and most importantly, a complete terminology has not been defined that describes the basic attributes associated with bird monitoring data. Categorizing the various survey methods (i.e. various point count methodologies, migration monitoring studies, banding stations protocols) into useable entities, which can be integrated through a variety of analytical tools, is essential.

Based on these challenges, several goals can be established. The first and foremost is to preserve and expand the bird monitoring dataset. Second, it is to make the information more accessible to people (including educators, policy makers, land managers, conservationists, and researchers). Finally, it is to integrate the bird monitoring dataset seamlessly with other online resources for use in multidisciplinary activities. The result will be that the contribution of the many monitoring projects to a larger network will create the "authoritative" source of information on birds and their environments.

Internet Technology Solution

There are Internet solutions to the challenges of organizing distributed and disparate data sets, but many of these are seriously flawed for the following reasons:

- There is only a very rudimentary "lingua franca," or common language, used by the numerous bird monitoring organizations for sharing information.

---

[1]A version of this paper was presented at the **Third International Partners in Flight Conference, March 20-24, 2002, Asilomar Conference Grounds, California.**

[2]Cornell Lab of Ornithology, 159 Sapsucker Woods Road, Ithaca, NY 14850. E-mail: stk2@cornell.edu

[3]Canadian Information System for the Environment, Hull, Quebec, Canada.

- There is no meta-catalog for organizing and accessing the bird monitoring information.
- There are only limited physical means to exchange and intermix data across disparate data formats and access policies.
- There are no means to determine the accuracy or authoritative nature of the bird data held within the distributed datasets.

Recent advances in several information technology sectors have successfully organized distributed data resources for collaborative computing. What follows is a brief overview of each of these advances, and how they can be integrated into a functional application to allow the bird monitoring community to organize access to their distributed data resources.

Data Grids

Advances in two distinct IT sectors provide the foundation for collaborative computing and information sharing at a global level. First, the ubiquitous availability of broadband networks such as the commodity Internet and Internet 2 provide the connectivity necessary to share information at a global level. Second, progress in distributed computing technologies, which have been available but on a smaller scale for many years, provide a means for collaborative computing across an integrated data environment. Combining broadband networks with distributed computing technologies has led to data grid computing strategies. These are based on a software infrastructure (eXtensible Markup Language (XML) requests bundled within Simple Object Access Protocols (SOAP) that link multiple computational resources via broadband networks. Data Grids provide solutions for decentralized projects that require a high level of computing, with data widely distributed across a network of users and resources. For distributed data collections, such as those for bird monitoring applications, the data grid model provides a single point of access for referencing data stored on multiple storage systems. The result is that from a single location, such as a web URL, all information stored within the data grid becomes available.

Metadata

The development of a metadata (information about data) standard for the bird monitoring community is central to making any vision of organizing the disparate bird data sources into a reality. In order to provide grid access to these distributed data resources common descriptors must be developed that associates the basic attributes of the disparate sources. By creating a relevant metadata schema the bird monitoring community will create a standard mechanism that describes the content, characteristics, condition and other qualities of the data. Creating this schema is not a trivial task, and is where the IT technologists interface with the bird monitoring experts.

The relationship between the metadata, and the data sources the metadata describe, supplies access points to data sources with similar content, even though the structure of the sources may be different. The most developed metadata standard is Dublin core (http://dublincore.org/). Developed by the Dublin Core Metadata Initiative, which promotes the adoption of interoperable metadata standards, Dublin core provides a common description for many basic attributes associated with publishing. It will be necessary for the Bird monitoring community to develop a specialized metadata vocabulary for describing its resources to enable more intelligent information discovery systems. Some progress has been made in this direction. For example, the Darwin core (http://speciesanalyst.net/docs/dwc/), developed by the University of Kansas Natural History Museum, describes the minimum set of attributes for search and retrieval of natural history collections and observation databases.

Data Handling and Resource Sharing

Middleware services are sets of distributed software that connect separate applications across the Internet and make resource sharing transparent. They act as the "glue" that integrates and organizes a network of servers, each brokering a specific set of data resources, into a single logical collection. Users can connect to any server in the network, and have full access to any data resource on it.

There are several client/server based middleware applications potentially useful for the bird monitoring community. These applications function within a federated server system in which each server manages a set of storage resources, and all servers respond to requests emanating from any of the servers on the network.

Species Analyst (http://tsadev.speciesanalyst.net), developed by the University of Kansas Natural History Museum, is a client/server application based on Microsoft Windows technology and uses a version of the Z39.50 server protocol developed by the Library of Congress. Its primary function is to search and retrieve information from distributed databases as if they were one. Species Analyst functions similar to any web server in that it listens to TCP/IP traffic. When a networked client makes a request, the Z39.50 server generates an SQL query to the database it supports, and obtains the requested information. The second genera-

tion of Species Analyst, DigiR, is currently under development. DiGiR (http://digir.sourceforge.net) obviates the need for a Z39.50 server by supporting web services using XML and SOAP. Species Analyst can support connections to a variety of database management systems such as Microsoft SQL Server, Oracle, Informix, Access, Paradox, DBase, FoxPro, Excel and even Text files.

Storage Resource Broker (SRB) (http://www.npaci.edu/dice/srb/), developed at the San Diego Supercomputer facility, is an across-platform application that provides a uniform interface across heterogeneous data resources. Users are provided with a set of operations to create, maintain, view, and search metadata objects within SRB. These metadata objects are organized in a meta-information cataloging system (MCAT) that can manage a large number and types of data sets of unlimited size. MCAT manages the attributes of all the collections within the grid's domain through a hierarchical metadata structure. When a request is made in SRB it is first routed to MCAT where the attributes of the requested information are determined. A query generator then embeds within an XML statement the appropriate query schema, relationships, and semantics and sends this to the correct SRB server. For example, if the requested information were stored in a database MCAT would generate an SQL query, but if the information were stored within a Unix file system then a Unix file system-specific query would be generated. Finally, a 'ticketing' system allows providers to designate which data can be made public and which data, e.g. on sensitive species, should be provided only to authorized users. Active development is underway to ensure that SRB has the appropriate application programmer's interface for as many different types of data handling processes as possible. SRB can access data in a variety of formats, held on many different file systems (Mac, Unix, Windows NT), platforms (disk or tape), and in varying formats (databases, text files, ADSM, object files).

Information Discovery and Presentation

Once the bird monitoring metadata attributes have been created, and a meta-catalog established, then the bird monitoring community can use data grid and resource sharing technologies to provide access to their data. Access to this information will provide a tremendous resource to assist educators, policy makers, land managers, conservationists, and researchers for ecoregional planning, and to make biological assessments for bird conservation. But simply providing access to the data is only the first step. Using the digital library approaches for information discovery and presentation, the bird monitoring data can be assimilated into a much broader set of data resources, which could provide meaningful visualizations, analysis packages, and other services. For example these applications could provide authentication and security for the collections, provide an organized data structure with easy navigation, provide tools for data analysis and visualization, and are made to be scalable to handle datasets of any number or any size.

What follows are several case studies that describe approaches which assimilate disparate data resources into meaningful representations. Once the technical considerations mentioned above are achieved, then the following is only a glimpse of what we firmly expect to be possible.

The Miistakis Institute (http://www.rockies.ca/) has developed a conservation planning portal focusing on the birds of the Northern Rocky Mountains. This portal, which is part of the Yellowstone to Yukon Conservation Initiative, allows conservation planners to access bird data from a distributed set of databases using the Species Analyst middleware application (http://www.rockies.ca/birds). These data are then superimposed on numerous GIS coverages and presented via the Internet on interactive maps using ESRI mapping products.

The Hayden Planetarium (http://www.haydenplanetarium.org/) at the American Museum of Natural History (AMNH) wanted to present a visualization of star and emission nebula evolution. To accomplish this involved integrating data resources, computational power, and manpower from three sites (AMNH, the National Center of Supercomputing Applications in Urbana Ill, and the San Diego Supercomputer Center). Using two supercomputers, over 1100 processors, and a total of 7 terabytes of data in 30,000 data files were used to render over 116,000 images. The final result was a 3.2-minute presentation showing the evolution of a star that can be viewed at the Hayden Planetarium. The three sites were networked via Internet 2, and all of the data was managed with Storage Resource Broker middleware (http://access.ncsa.uiuc.edu/Releases/02Releases/03.07.02_San_Diego_.html).

The Cornell Lab of Ornithology (http://birds.cornell.edu) has developed interactive mapping tools which access mapping information from distributed resources. For example, eBird (http://www.ebird.org/beta/MyEBird), a joint project with Audubon, integrates aerial photographs and topographic maps from Microsoft's TerraServer (http://terraserver.microsoft.com) with information stored in databases at Cornell, to assist participants in pinpointing the locations where they have made bird observations. Both web services and application programmer interfaces are provided to

allow access to the images and data stored within the TerraServer database.

Conclusion

The technology is now available to interconnect all of the bird monitoring data assets through a single accession point while still providing local control of the individual data resources. Using data grid strategies, information access over a widely distributed network of resources and users in the bird monitoring community can be accomplished.

An essential ingredient to make this successful is a metadata information-tagging schema that provides the common protocol that links all of the disparate information sources. While some progress has been made in developing this metadata schema, simply extending the Dublin Core metadata structure is not sufficient. It is recommended that a concerted effort by both biologists interested in monitoring issues, and technologists who understand digital libraries and their technology be made. For example, there are almost as many bird count protocols as there are data resources. Developing a hierarchical metadata structure that allows a stepped integration of data from these count types is necessary.

Most, but not all, of the bird monitoring data is stored "electronically." But this data is stored in a variety of formats; some in databases, some as text or html files, others as rich media (images and video), and some of this data may be unreliable or erroneous. Providing access to this information with the appropriate considerations of data security and reliability is of paramount importance. With an ever-increasing variety of bird monitoring information available it is essential to expand beyond database queries the middleware applications used in the bird monitoring network data grid scheme.

Once the underlying architecture of interlinked data resources is achieved, emphasis can shift to expanding the digital library services available. These services can present the data in meaningful visualizations, and analysis packages for bird conservation, research, and education. Access management strategies can be implemented to ensure that access to sensitive information is restricted. Analytical nodes that use all of the grid's data resources can provide tools that enable users to understand the trends and dynamics of bird populations. Comprehensive educational pages that integrate the analytical results for a particular species with videos, sounds, maps could be developed. Finally, collaborative computing projects that use the expertise of several partners can use the large volume of data available on the grid, to generate visualizations and analysis for research, education, and bird conservation.

Literature Cited

Brackett, David. This volume. **Bird conservation as a flagship for global diversity conservation.**