INTRODUCTION

The past decade has brought a growing awareness of the information crisis facing the ecological sciences. A recent editorial in Nature (Nature 1999) on the inability of the current information management infrastructure to absorb the sheer quantity of biological data to be produced by the next generation of ecological monitoring initiatives indicated that these concerns are in the forefront. Among our most immediate concerns are:

Access to primary data. One of the most pressing problems facing the biological sciences is the long-term survival of research data. Future research projects are likely to show an increased reliance on archived datasets as primary empirical data sources, because of the complexity and increasingly interdisciplinary nature of ecological research (especially that concerning urban environments). Only recently have we begun to develop systematic solutions for the electronic preservation and publishing of primary data, such as the Oak Ridge National Lab (Olson et al. 1996, Olson and McCord 1998), San Diego Supercomputing Center (SDSC) (http://ceed.sdsc.edu), and the Long-Term Ecological Research (LTER) network (Porter and Callahan 1994, Stafford et al. 1994). Although these efforts enhance the survivability of new data from certain sources, there is a tremendous backlog of existing (legacy) datasets for which the costs of archive preparation and management are prohibitive.

Locating and identifying relevant information. Growing quantities of online data necessitate the indexing and documentation of information. Although simple text-based search engines are freely available, they are ineffective for indexing datasets because they do not discriminate among categories or contexts of information. An efficient system to record metadata will greatly enhance the utility of the preserved data. The term “metadata” is now widely defined as information that identifies and describes a given set of data. Several efforts have been made to publish either in final or draft form, content standards for metadata (Federal Geographic Data Committee http://www.fgdc.gov; Michener et al. 1997; National Biodiversity Information Infrastructure 1998 http://www.nbii.gov/).

Diverse and dynamic state of information storage formats. Storage and man-

<table>
<thead>
<tr>
<th>Acronyms used in this proposal.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASU</td>
</tr>
<tr>
<td>CAP LTER</td>
</tr>
<tr>
<td>CD ROM</td>
</tr>
<tr>
<td>CES</td>
</tr>
<tr>
<td>CIPE</td>
</tr>
<tr>
<td>DTD</td>
</tr>
<tr>
<td>EDA</td>
</tr>
<tr>
<td>ESRI</td>
</tr>
<tr>
<td>FGDC</td>
</tr>
<tr>
<td>GIS</td>
</tr>
<tr>
<td>HTML</td>
</tr>
<tr>
<td>K-12</td>
</tr>
<tr>
<td>KU-NHM</td>
</tr>
<tr>
<td>LTER</td>
</tr>
<tr>
<td>NACSE</td>
</tr>
<tr>
<td>NCEAS</td>
</tr>
<tr>
<td>NBII</td>
</tr>
<tr>
<td>NSDI</td>
</tr>
<tr>
<td>NSF</td>
</tr>
<tr>
<td>RAD</td>
</tr>
<tr>
<td>SDE</td>
</tr>
<tr>
<td>SDS</td>
</tr>
<tr>
<td>SQL</td>
</tr>
<tr>
<td>USAID</td>
</tr>
<tr>
<td>USGS</td>
</tr>
<tr>
<td>XML</td>
</tr>
<tr>
<td>XSL</td>
</tr>
</tbody>
</table>
agement of ecological data are (and will likely continue to be) done at a local institutional level, close to the individuals who generated the data. Significant advances have been made in the last decade to achieve Internet connectivity to major repositories of ecological data. However, extant differences in data storage schemas, file formats, and communication protocols require further development of enabling technologies that permit query and retrieval from heterogeneous systems in remote locations.

Targeting user audiences. Although online data availability is a significant step forward, usage often falls short of expectations due to the learning curve required to use archived data. Electronic applications of primary and synthesized data need the product-oriented perspective—the ability to target audiences—of conventional publication media.

In ideal circumstances, information technology should provide a hierarchical structure in which observations by individual scientists at discrete points in space and time can be stored, documented, retrieved, interpreted as information and, ultimately, contribute to our general knowledge about the world around us (Figure 1). If we are to achieve this infrastructure, then we must carefully plan our investments in developing the key components of technology and organization.

![Figure 1. Hierarchical model of data infrastructure.](image)

**PROJECT GOALS**

We propose to address these needs by creating an infrastructure based on tools, data archives, and applications to deliver ecological data products in more useful ways. To realize this goal, we will draw upon the data resources available at Arizona State University (ASU) and the Central Arizona – Phoenix Long-Term Ecological Research (CAP LTER) project. Membership in the LTER network creates an ideal focal point for developing broader data archiving and access infrastructure that supports non-LTER data produced at ASU and by local, state, and federal agency partners in the environmental field. The *Networking our Research Legacy* project will focus on three areas:

- Developing key technological components to better document research datasets and make them more accessible. Using eXtensible Markup Language (XML) as the underlying medium for information exchange, we will develop a series of software tools that: 1) assist in creating and managing metadata; 2) export and exchange metadata (and optionally the primary data themselves) in XML format; and 3) use these metadata to automatically generate data preview and query applications so that a more generic query interface can be maintained to access many datasets from heterogeneous sources.

- Developing data delivery systems that bridge the gap between data archives and users by targeting three classes of user groups: 1) the ecological research community as represented by ASU students, faculty, research scientists, and the broader LTER network; 2) the K-12
Systematics and Collections Databases at ASU.

ASU's Entomology, Herpetology, and Ichthyology Collections. The Hasbrouck Insect Collection, which contains over 2,500,000 specimens, is the primary reference collection for regional insects. The catalog database was recently upgraded through CAP LTER funds. The herpetology and ichthyology collections, consisting of over 28,000 and 16,000 specimens, respectively, are extremely valuable due to their Southwestern emphasis and the endangered or threatened status of many species. With support from a LTER supplemental grant, electronic inventories for these collections were merged and migrated to a relational database server to enable Web access and integration with CAP LTER research data. The CAP LTER data manager is working with the collections' curators to finalize the Web-access interface.

ASU's Vascular Plant Herbarium is the second largest in the arid Southwest, with over 210,000 mounted specimens. Collections from various floristic treatments in Arizona and Mexico make this herbarium a unique resource, now recognized as one of only 105 resource herbaria in the U.S. Herbarium research centers primarily on Southwestern flora and biosystematics of Cactaceae, Compositae, and Myrtaceae. A static species list is available through ASU's Web site; providing searchable access to this database will be a product of CAP LTER activities.

ASU's Lichen Herbarium is the largest in the Southwest, with over 63,000 accessioned specimens. An additional 30,000 specimens are being evaluated in the context of the NSF-funded Sonoran Desert Lichen Flora Project, involving over 50 professional lichenologists from around the world. Taxonomic treatments of Sonoran lichen groups are integrated into an online world key of lichens (http://www.botanik.biologie.uni-muenchen.de/botsamml/lias/lias.html). With assistance from the Northwest Alliance for Scientific Computing and Engineering in Corvallis, Oregon, the herbarium's database of over 42,000 records is now fully searchable at http://mgd.NACSE.ORG/cgi-bin/qml2.0/arizona/arizonaHerb.qml. A taxonomic checklist of species names and synonyms is being accumulated, which consists of over 17,000 entries and is worldwide in scope. A comprehensive database of literature on lichens is available at http://www.tøyen.uio.no/botanisk/bot-mus/lav/sok_rll.html.

educational community; and 3) resource managers and conservation ecologists.

- Enhancing existing ASU data resources for these two tasks by: 1) enabling search access to core databases via national search protocols; 2) implementing geographic indexes and software to facilitate spatial query of legacy databases that lack explicit geo-referencing; and 3) applying cross-referencing indices to better integrate the core of referencing and cataloging databases. These steps will leverage the value of each resource through cross-referencing and enable national access through universal search protocols used by library search engines.

Significance

The significance of this project is two-fold. First, the project addresses several critical and well-documented needs in the management of ecological data with a modular set of technical solutions designed to be transferable and forward-compatible. Second, the integrated approach taken by this project in seeking to target audiences and potential users of ecological data will leverage the value of prior investments in research grants leading to the accumulation of those data.

ASU ECOLOGICAL DATA RESOURCES

ASU has accumulated a significant set of ecological data resources. These include taxonomic inventories of organisms and their occurrences in the study area; specimen collections catalogs for entomology, herpetology, mammalogy, ichthyology, and plant biology; and bibliographic databases of published and unpublished literature relevant to the research, taxa, and collections referenced above. In addition to these synthesized databases, smaller individual datasets resulting from discrete research projects are rapidly accumulating through ongoing activities such as the CAP LTER.

Systematics and Collections Databases

Biological databases of primary relevance to this project (see sidebar Systematics and Collections Databases) are in the Departments of
Biology and Plant Biology. The Biology Department houses cataloged collections of insects, amphibians, reptiles, and fish, as well as some birds, mammals and shells, while Plant Biology houses two major formal plant repositories: the Vascular Plant Herbarium and the Lichen Herbarium. Although some of these databases either are, or will soon be, accessible via Web-based query tools, each system is inherently proprietary and currently not well suited for integration into a broader network of scientific and museum inquiry.

**Bibliographic Databases**

CAP LTER's bibliographic database contains over 2000 references to central Arizona-Phoenix ecological research in a relational database with a Web-based user interface. The interface supports search, display, and saving of bibliographic selections. A modified version of this format is now being developed for use as the LTER network exchange standard for bibliographic data. In addition to the CAP LTER database, extensive bibliographic holdings are contained in the Lichen Herbarium database and in a database compiled by Biology Professor Minckley. ASU recently collaborated with the University of Arizona on a proposal to adopt the CAP LTER bibliography database format into a statewide system to inventory cultural and environmental literature from the state's four major museum libraries.

**Metadata Catalog**

Metadata are currently compiled for CAP LTER datasets in a relational database. The content of this database includes all relevant elements defined in the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata. Additional elements were included to accommodate information specified in the Michener et al. (1997) paper on ecological metadata. This data catalog presently contains descriptions of some 50 datasets and offers rudimentary search capabilities via a proprietary Web interface on the CAP LTER Web site. Where available, live links to downloadable files are provided. The metadata catalog is an essential resource as it serves as the gateway between major search applications that help users locate primary datasets and the applications that provide online access to those resources.

**Primary Research Databases**

The CAP LTER project is expected to contribute many primary datasets to be archived at ASU and indexed by this system. Although the proposed project will not fund the creation, documentation, and archiving of primary research data, this work will nevertheless benefit greatly from the project's technological developments, which far exceed the scope of software development that can be supported by any single research program. In addition to the CAP LTER projects, ASU has a long history of ecological research carried out with funding from NSF and others. The current list of active National Science Foundation (NSF) grants for ASU lists over 20 projects related to lab or field biology that can contribute to the existing body of research data at ASU. One notable example is the 20-year dataset produced by Grimm and Fisher's research in Sycamore Creek, a NSF Long-Term Research in Environmental Biology project that will be among the first non-LTER datasets to be published through this data infrastructure. In an environment heavily influenced by human activity, datasets covering a wide range of environmental and cultural parameters are necessary to provide meaningful contexts for biological data and research. Successful use of these data depends on the accuracy and efficiency with which acquisitions can be processed and made accessible to researchers. In many cases, information providers do not have the resources to compile full metadata nor provide support for data use. CAP LTER and other ASU projects have acquired through data-sharing arrangements substantial datasets from various partners (Table 1). Delivery of these datasets to ASU researchers and (where appropriate) non-ASU users will be greatly facilitated by software that can reduce the time and expertise required to ac-
cess and process shared data.

In this project we will develop tools and applications to enhance existing abilities to manage, access, and make productive use of databases such as those described above. A primary need for this expansion lies in the anticipated pace of accumulation of new data either directly from the CAP LTER project or from future research efforts that follow its lead. Although it is expected that such projects will, as does LTER, provide funding for long-term data management, these are not usually adequate to support the development of innovative technologies that enable us to deliver data products in more useful ways.

**PROJECT TASKS**

**Tools: Develop Kernel Technologies for Metadata and Data Processing**

It is widely recognized that metadata is the cornerstone to advances in information management. To advance online data delivery systems to use more automated procedures for accessing and processing data, it is necessary to encode metadata in formats that are portable and machine-readable. eXtensible Markup Language (XML) has received much attention as a technology for encoding structured information (such as data or metadata) for exchange across the Internet and among applications (Jones 1998). Like HTML, XML is a subset of Structured Generalized Markup Language that uses inline tags to identify and attribute content in a text document. As with HTML, XML is a public, nonproprietary language maintained by the international W3 Consortium. One of the most significant features of XML is the ability to define a schema, known as a Document Type Declaration (DTD), that can be used as an external reference against which a document may be validated to enforce strict compliance with a given schema. This is a critical property that makes XML suitable for encoding machine-readable information.

At the kernel level, the project will develop software tools that make use of XML-formatted metadata to automate many of the costly tasks associated with acquiring and disseminating data products (Figure 2). Although their development will facilitate the building of a networked data infrastructure at ASU, the ultimate goal will be to maximize the transferability of these tools through the use of open technical standards such as XML and Java.

For this task to proceed, it must be designed around an appropriate standard for ecological metadata. The most mature and appropriate current standards are the Federal Geographic Data Committee (FGDC) content standard for geo-spatial metadata and the National Biodiversity Information Infrastructure (NBII) biological extensions to this standard that govern the represen-
Ecological Analysis and Synthesis (NCEAS)’s recent project to develop a series of XML DTD’s based on the Michener et al. (1997) metadata standard (Nottrott et al. 1999), and the Northwest Alliance for Computational Science and Engineering (NACSE)’s Java based tools for extracting database structure via Java Data Base Connectivity protocol. Because standards are evolving, synchronization with the larger community of scientists responsible for their definition is essential. To ensure such synchronization, we will host a small workshop shortly after completing the draft metadata schema. Participants will be sought from several institutions that clearly exhibit leadership roles in the rapidly moving world of metadata standards and associated applications. These include NCEAS, FGDC, NBII, the LTER network office, University of Kansas Natural History Museum (KU-NHM), NACSE, and SDSC. The goal of the workshop will be to solicit input before software development begins and communicate project activities to those engaged in similar research.

Creating metadata

To facilitate acquisition of datasets into the infrastructure, we propose to develop and adapt tools for harvesting metadata from legacy databases. With support from NSF, NCEAS has developed a prototype editor for producing metadata in XML format compatible with the Michener et al. (1997) metadata standard. The project will build upon this work by developing two integrated components based on the “wizard” model exhibited in many modern computer programs, such as tax preparation tools. The first is a set of functions to reverse-engineer table, column, and relationship information and, for Geographic Information Systems (GIS) and remote sensing data, spatial reference information. The second component will be a user-friendly interface for harvesting information that cannot be reverse-engineered. In most cases, the most accurate source of this information is the scientist who produced the dataset. However, as familiar as most scientists are with their own data, few are familiar with the complex metadata standards required for documentation. Therefore, an application that queries the user for information will simplify and streamline the process of gathering metadata information, allowing data producers to shoulder the burden of preparing datasets for archiving and publication. The program output will be a complete, or nearly complete, XML document that can receive final edits from a database manager using the simpler editing application developed by NCEAS.
Reporting and displaying metadata

The keystone to developing user-oriented applications for data products from a central data server is to provide accurate metadata about the delivered data. In many cases, this may be a subset created by a query from a much larger, more complex database. This project will develop reporting software to evaluate a query to a database, examine the metadata for that database, and then return, along with the requested data, an XML-metadata report custom tailored for the received data file. Only metadata for the retrieved fields or records will be included, resulting in a more accurate metadata document upon which client applications will be better able to make processing decisions.

XML tags serve only to indicate the content of a document. XML documents contain no instructions on how the information is to be displayed, because many applications of XML do not involve direct display or printing of the information. However, an XML document can be paired with one or more style sheets written in eXtensible Stylesheet Language (XSL), which determines how elements are to be displayed. In this project, XSL style sheets will be developed so that a single metadata document encoded in XML may be displayed or printed in several alternate formats such as FGDC-indented format.

Generating interfaces for data query

Enabling access to datasets stored in relational database systems typically involves developing a user-friendly query interface that accepts parameters, formulates a Structured Query Language (SQL) query statement, and returns the result in a usable file such as Excel or text. Although Rapid Application Development (RAD) tools make this process relatively fast, it still involves creating code that is essentially proprietary to the dataset. For large datasets, the relative gain for this effort is high, but with the smaller datasets typical of many research projects, it is difficult to justify the effort for producing a user-friendly forms interface. To reduce the cost of enabling access to research datasets, the project will develop a Java-based query interface that reads the metadata description of a dataset and uses that information to open a connection to the dataset and generate an on-the-fly query form that could provide a user with options for previewing or querying the dataset. Stored views of the data, documented in the metadata, would predefine the most common joins among tables, eliminating the need to specify these in the actual query application. Because most ecological datasets have some degree of spatial organization, this application will need to support GIS data types (points, lines, polygons, imagery) in these preview and query functions. Some of the most basic functionality described here has been developed by the Northwest Alliance for Computational Science and Engineering (NACSE) in a tool called Query Designer. This product will build upon this effort by drawing upon metadata from an XML document, rather than reverse-engineering minimal information from the data source, and by adding more extensive preview capabilities.

Exporting and importing data

In some instances, relying on native protocols for importing or exporting data is not the best approach for data exchange. Datasets based on multiple, related tables are difficult to migrate because of the complex pattern of dependencies between tables that must be maintained during the exchange process to avoid key violations. Because XML documents can be enforced to adhere to a schema, they provide an ideal medium for exchange of primary data in cases where programming in native protocols such as SQL can be laborious. To facilitate use of XML to exchange data as well as metadata, two components are required—one at each end of the exchange process. At the source end, an application is needed to read a dataset’s metadata and generate on-the-fly a DTD that matches the data schema. Records are then extracted and delivered in a conforming
XML document. At the destination end, an application parses the XML document and then, with the knowledge gained from the accompanying metadata, generates the necessary commands to insert the data into the target, such as an SQL database. A prototype of the receiving portion of this application was developed for inserting bibliographic data into the CAP LTER database. Expanding this prototype with the capacity to generate the appropriate processing code at run-time by reference to XML-encoded metadata will create a powerful instrument for data exchange. Development of XML data exchange will have significant effects on research projects relying on remote entry of data from volunteers or K-12 students, as it will eliminate the need for managing open connections, security and quality control through complicated, proprietary entry applications. Complex data can be entered into a local form then submitted over the Internet as a single document, processed for quality control, and entered at the server using the above tools.

**Applications: Develop Targeted Data Delivery Systems**

The tools produced by the above activities will be applied in the applications stage of the project to develop data delivery solutions that can be directed toward well-defined user groups. The function of these applications is to hide the details of data storage, query, and retrieval to allow the user to concentrate on the information they want, rather than the complexities of its diverse sources. Attempts to articulate our efforts with programs having existing infrastructure for reaching target audiences will help leverage results. We will approach the task of application development in a modular manner, by combining tools developed in this project with other, established solutions to build open-ended applications which will be capable of delivering data from a diversity of sources and easily modified to accommodate different targeted groups (Figure 3).

We will demonstrate the effectiveness of this approach by developing a prototype series of applications that use common tools to deliver data from a common suite of data sources (bird, arthropod and vegetation monitoring) to three different audiences. For the purposes of this project, we identify three broad kinds of audiences: research, education and management. In each case, Internet-based technologies such as Java and map server software will be used to provide a visual interface through which query parameters are solicited from the user and results are displayed graphically. Although the underlying data and programming code will be similar, applications for each target group will differ in the degree of choice, data precision, and units of presentation. In principle, these applications would be easily extensible to virtually any dataset that is documented using the metadata standards developed for this project.

In addition to this prototype application, we will also seek to apply the infrastructure tools created here to develop articulations between existing programs within each audience group and the data resources accessible at ASU.

**Academic research**

Research users represent the primary target audience to be addressed in this project. The very nature of the scientific process renders researcher needs less predictable than other groups, requiring that applications support flexibility in search parameters and output content. These needs
will be addressed in our prototype application by providing intelligent preview and query features that ease access to the primary data. Preview capabilities would allow a user to retrieve Exploratory Data Analysis (EDA) summaries such as quartile breakdowns, histograms, or box-and-whisker plots. Query functions would assist the user in creating the necessary Structured Query Language syntax to query the dataset and retrieve a file in some chosen format.

The importance of collections databases being accessible on-line has been recognized by the establishment of national clearinghouses. A significant contribution of this project to academic research will be the increased availability of information gained from integration with existing national search protocols based on the Z39.50 query standard. As part of this phase of the project, we will install a local Z39.50 client to allow development of query interfaces that can take advantage of the spatial search enhancements developed in this project. Improving spatial analysis of collections data will turn these databases into a powerful tool for understanding species distributions as well as taxonomic variability and analysis of ecological communities.

By adopting XML as the underlying standard for data and metadata handling, information resources will be better positioned for potential integration into the next generation of Internet-based research applications. One of the most significant applications for datasets resulting from discrete research projects is modeling. The CAP LTER research design calls for extensive development of landscape models to describe and explain ecological processes in the urban environment. The XML tools developed in this project will be applied to enhance the delivery of sampling and core monitoring data directly into the applications currently under development by the CAP LTER modeling team. This will not only facilitate data access to modelers, but will open the possibility for building models based on live, or near-live data feeds from core monitoring activities. The feasibility of this approach has been demonstrated by the Internet-based species modeling application under development at the University of Kansas Natural History Museum and San Diego Supercomputing Center (Kaiser 1999). Developing seamless access to primary data from landscape modeling applications will directly benefit this major component of CAP LTER research.

**K-12 education**

In developing our prototype data presentation applications for educational audiences, emphasis will be placed less on query options and more on summarizing data according to units that can address the interests of site visitors. One application goal will be to develop a spatial display of ecological monitoring data using reference layers that would emphasize school locations and city boundaries, allowing students to visualize ecological data from the varying perspectives of home, school, or neighborhood. Navigational features would include the ability to zoom to individual street addresses or other common landmarks.

A primary object of this portion of the project will be to develop articulations with existing programs designed to interface between ecological research and the educational and public communities. The Ecology Explorers program, part of the CAP LTER project, is combining research and education in a unique manner that will benefit from new database technologies. Students from throughout the metropolitan Phoenix area are collecting data on a variety of different ecological communities to test hypotheses about the impact of urbanization on ecological systems. For example, students are measuring plant diversity, collecting data about arthropods that captured in pitfall traps, conducting bird censuses and sampling communities of seed eating beetles. They can access protocols explaining how the data are to be collected, enter data, and look at results of the study through the Ecology Explorers page on the CAP-LTER Web site. To fully utilize the potential of this project students will need to be able to access much more than the portion of the data that they collect. Developing solutions for students to integrate their own data with that of larger research projects expands the range of options open to them for analytic exercises and is an
explicit goal for the Ecology Explorers program. The data infrastructure developed here will enable students to access the broad wealth of data collected by mainstream research, downgrading its complexity to make it compatible with the units and scales at which their own data are collected. One example of this integration will be to incorporate an existing feature on the Ecology Explorers Web site—an interactive species identification key—as an interface to searching the core taxonomic databases, thus providing an entry into a rich set of information that would be otherwise difficult to navigate using the conventional search criteria and interfaces created for academic users.

Another program that would benefit from infrastructure is the Center for Image Processing and Education (CIPE), a NSF-funded center specializing in electronic classroom materials that teach information technology. Presently, CIPE solicits data from real world research, synthesizes them for teaching purposes, and distributes them as part of a CD-ROM package. This project will work with CIPE to explore solutions for making data available to CIPE tutorials over the Internet. The advantage gained will be the ability for teaching applications to make use of more current, or even real-time, data that is not possible using CD-ROM distribution.

Resource management and conservation ecology

Data products generated through academic research are of potential value to resource management and conservation decision-making. However, too often there is little exchange of information between academic researchers and natural resource managers. A recent ASU consulting project to determine priorities for purchasing land for preservation of biological diversity in the Cave Creek Wash area in Northern Phoenix, found that data on biological diversity in the area was scattered or absent and it had never been organized in a geographical context. Tools developed through this project have the potential to increase access to databases that are essential for management-oriented research.

Management queries to information systems are commonly expressed in spatial terms—that is, information is sought first for a specific bounded area based on legal or administrative criteria. We propose to provide our prototype application with an interface that supports sophisticated expression of spatial queries via online, graphic tools and yields results using visual output. One form this interface might take is a Web-based sensitivity map that interactively summarizes selected data according to spatial units that have particular relevance to management activities such as planning units, land ownership boundaries, or census blocks. Easily zoomed and panned, the application would provide potentially useful management information without actually releasing primary data, thus avoiding security or intellectual rights issues that might preclude such efforts.

**Databases: Enhancing ASU’s Data Resources**

With funding from this grant, we will implement several improvements to the core set of reference databases that leverage their value and increase their compatibility with the range of technological solutions implemented in this project (Figure 4).

**Establish a node for national clearinghouses**

To enable national access to these resources, the Center for Environmental Studies (CES) will establish itself as a contributing node to three closely related clearinghouse networks that provide remote search capabilities for environmental information: the National Spatial Data Infrastructure (NSDI) clearinghouse for digital geospatial data (http://www.fgdc.gov), the National Biodiversity Information Infrastructure (NBII) clearinghouse for biological data (http://www.nbii.edu), and the University of Kansas (KU-NHM) Experimental Biodiversity Information System (http://chipotle.nhm.ukans.edu). Each system uses the Z39.50 protocol for handling communications between multiple, contributing nodes and a central search client which harvests and merges
results into a single output document. Databases are made accessible to the remote system by installing a Z-compatible server and developing a schema that maps the content of the database to standard elements (such as author, title, date) used in the Z39.50 protocol. Funding has been received from an LTER supplemental grant to install Z39.50 services at CES and to integrate the collections catalogs with the KU-NHM network. Funding through this grant will allow us to extend this protocol to the metadata catalog and bibliographic databases and to the NSDI and NBII networks.

**Develop a spatial search interface**

Many legacy datasets, especially collections and bibliographic databases, are not geo-referenced, making spatial queries difficult. The cost of geo-referencing (that is, digitizing a spatial shape for each record entry) large legacy databases can be prohibitive. In many existing databases, spatial references exist in their use of place names in keyword lists and location descriptions that presently go unused by most spatial search engines. For example, a search on “Maricopa County” would fail to return records that were keyworded only by “Phoenix” or “Scottsdale”. We will build a tool to tap the information in these keywords by loading a Spatial Database Engine (SDE) database with records from common place name directories such as the United States Geographic Survey Geographic Name Information System. We will then develop routines in C (the language supported by SDE’s development tools) that can return a list of records whose attributes match a submitted text string and, inversely, can return a list of keywords for shapes that spatially match with a submitted shape. With these simple tools, we can enhance the power of existing search systems by allowing spatial searches to be run from a text-based interface against databases that have not been explicitly geo-referenced by capturing the search term and searching the place name directory for a match. If a match is found, we can then locate other spatially convergent shapes and return their attributes as an extended set of terms for searching the target database. For interfaces that do permit graphic expression of a search boundary, this tool will still enhance their ability to work against databases that are not geo-referenced.

**Apply cross-referencing indices**

The relationships among the core datasets (systematics and collections, bibliography, metadata catalog) will be made more explicit through cross-referencing tables and indexing. These improvements will support development of applications that integrate content from one or more of these systems, making it possible, for example, for a researcher to locate a dataset and quickly access related publications or taxonomies.
Partnerships: Building an Ecological Information Network for Arizona

The final project task seeks to ensure the long-term sustainability of the infrastructure by encouraging a more formal data sharing partnership among agencies (both producers and consumers of data) in Arizona. A model for this partnership is the Oregon Coalition of Interdisciplinary Databases (http://www.nacse.org/ocid/) that consists of partnerships among agencies maintaining databases not unlike the list of resources targeted by this project (they encompass one LTER site, academic species lists, collections catalogs, and more). The infrastructure would enable these databases to remain relatively autonomous in their management, yet integrated via the technical metadata tools and the cross-referencing of our core databases. To promote this partnership, CES will organize meetings with local agencies to explore potential advantages to them in the infrastructure developed in this project and to encourage expanding the network to include shared, online data resources from their agencies. In addition to the agencies listed in Table 1 that have already shared data with ASU, potential invitees would include Arizona Game and Fish, the State Climatologist’s Office, city, and federal agencies. The goal for these meetings will be to develop long-term strategies for connectivity and data sharing using tools developed here.

PROJECT TIMETABLE AND DELIVERABLES

Figure 5 plots the approximate timetable for project activities by task and project year. In Year 1, database activities will begin. Content standards for the metadata format will be selected and the XML DTD’s will be developed. Profiles to map data fields from the relational databases and the Z39.50 elements will be developed. In the latter half of Year 1, a workshop on metadata standards will be held and programming will begin on the first of the XML tools, the metadata entry wizard.

In Year 2, programming XML tools will continue. Connectivity to the national search networks will be completed, as will the spatial search processor. Applications projects will commence in Year 2, beginning with the development of research-targeted applications followed by educational interfaces.

Activities in the final year will be spent testing tools and developing end-user applications for ecological data. Research and educational applications will continue to be refined with feedback from test users. The planning meeting for developing community partners in an ecological infor-
Information network for Arizona will be held early in Year 3, with feedback from that meeting refining the goals for integrating research data in management applications.

**Y2k Statement**

Public release documents regarding Year 2000 status have been obtained from vendors for primary products proposed for use in this project. These include Intel Pentium II microprocessor and associated motherboards and bios, Microsoft SQL Server 7.0, NT Server 4.0 service pack 5 and Visual Studio 6.0, Environmental Systems Research Institute’s Spatial Database Engine 3.0.2 and MapObjects 1.2, and Symantec Visual Café Java Development package 3.0. None of the proposed products indicate functionality issues related to two-digit years or other Year 2000 phenomena.

**Policy Regarding Intellectual Property Rights**

Individuals or institutions identified in the accompanying metadata descriptions of resources will retain intellectual property rights over datasets accessed via this system. Access to all data resources will be moderated through a free user-registration system and governed by the Data Access Policy developed for the CAP LTER Web site (http://caplter.asu.edu/data). All software produced by this project will be made public domain with code and binaries freely available via the Internet.

**SENIOR PERSONNEL**

*Peter McCartney* is an Assistant Research Professor at ASU and has served as Information Manager for the CAP LTER project since 1997. Additionally, McCartney has been involved in data management, GIS training and research at ASU since 1992. As information manager for the Archaeological Research Institute, he has developed online publications of archaeological research datasets with metadata and directed several large collaborative database projects serving management and academic needs. He has developed relational database systems for managing collections, numbering in the hundreds of thousands of items, for both the Biology Museum collections and the Archaeological Research Institute. McCartney will serve as co-project director in charge of research, overseeing the design and development of software and applications.

*Charles Redman* is Director of the Center for Environmental Studies and the Archaeological Research Institute and is co-Director (with Grimm) of the CAP LTER project. As director of CES and administrative co-director of this project, Redman assumes responsibility for the long-term sustainability of the information network produced here and the administrative needs of this project. He will also coordinate meetings with community partners to develop interest in participating in an information network based on the infrastructure developed.

*Corinna Gries* is an Assistant Research Professor at ASU in the Department of Plant Biology and serves as database manager for the Lichen Herbarium. Gries is an ecologist with database and programming skills, who has been involved in interpreting science to the broader public. She is co-PI of the recently completed Sonoran Desert Lichen Flora project, has conducted field projects for the scientific collection of data, has developed collections and research databases, and has been involved in teaching and facilitating inquiry-based learning to K-12 and university students. Gries will serve as project manager and coordinate the process of bringing research databases online, testing the tools developed for collecting metadata, and providing data via the Web to a broader public of different user types. She will integrate databases into the networked system, including training and supervising students in metadata compilation, data conversion or importing, and track project progress.

*Tim Craig* is an Associate Professor of Life Sciences and a participating scientist in the CAP
McCartney - 1026169 Networking our Research Legacy

LTER project. His research background is in evolution and speciation, interactions between trophic levels, and developing programs involving K-12 students and teachers in research. His primary role in this project will be to oversee the application development activities relating to educational and management audiences.

Nancy Grimm is a Professor of Biology and co-Director (with Redman) of the CAP LTER project. Her expertise is in bio-geochemistry, water quality, and ecosystem ecology, and she has experience with long-term ecological data both through the LTER program and as co-Director of the 20-year Sycamore Creek project. Her primary role in this project will be identification and recruitment of databases from the ecological research and regulatory communities for inclusion in the study.

Other senior personnel

Leslie R. Landrum is Herbarium Curator in the Department of Plant Biology. He has research interest in the flora of Arizona and South America and is co-editor of the Journal of the Arizona-Nevada Academy of Science, flora issues. He is responsible for the technical design and management of the vascular plant database. His role in the project will be to assist in the integration of the vascular plant database and its taxonomic lists with the core databases.

Donald J. Pinkava is Professor of Plant Biology and a participating scientist in the CAP LTER project. His expertise covers plant bio-systematics and the vascular plants of the Southwestern U.S. and Mexico. His role in this project will be to assist in the integration of plant taxonomic databases and the end user interfaces for accessing them.

Two staff positions will be created for additional personnel. A senior database programmer will carry out all technical programming related to XML tools, systems software, and application development. A Web developer/graphics designer will be responsible for developing the functionality and look/feel of application interfaces. Both of these positions will be filled from a national search.

RESULTS OF PRIOR SUPPORT

Co-PI McCartney, with George Cowgill and Charles Redman, NSF SBR-9816263. $99,508.000. 1/1/99-6/30/00. Ensuring the Long-Term Viability of the Teotihuacan Mapping Project Database. The central Mexican city of Teotihuacan represents a pinnacle of prehistoric Mesoamerican civilization and has been a focus of archaeological investigation for more than a century. The Teotihuacan Mapping Project, begun in the 1960s, divided the 20km² ancient city into 5500 tracts and has described each tract in terms of more than 400 variables, including observations made during the survey and from ongoing studies of collections of approximately 1 million potsherds and other artifacts. Although data were computerized early on, much information potential of this dataset remains untapped. This project, which has just begun, will update and preserve the files and, in collaboration with the Archeological Research Institute, ensure their long-term integrity and broad accessibility, creating software to enable remote users to query the database through the Internet. The primary data contribution will be to create the extensive metadata documentation of the Mapping Project, providing a prototype for other archaeological projects. These enhancements will facilitate use of the data by project researchers and make the data available for the first time to others. In addition, a book will illuminate salient patterns in the data.

Co-PI: Nancy B. Grimm and Charles Redman, NSF DEB-9714833. $4,274,940 + 11 supplements. 11/1/97-10/31/04. Central Arizona - Phoenix Long-Term Ecological Research Project. The CAP LTER joins 21 other national sites charged with monitoring and assessing long-term ecological change. Objectives of this long-term study of the Phoenix metropolitan area and its expansion zone are to: 1) generate and test general ecological theory in an urban environment;
2) enhance understanding of the ecology of cities; 3) identify feedbacks between ecological and socioeconomic factors; and 4) involve K-12 students in scientific discovery. First-year activities centered on acquiring information to set up a rational, spatially based monitoring program. The project acquired data to: 1) develop a sense of the overall structure of the CAP study area, including historic spatial patterns; 2) define patch topology and long-term monitoring strategies; and 3) construct initial materials budgets for the entire valley’s hydrological system. Modeling has begun in concert with data synthesis and new data acquisition; 20 pilot projects were completed in the first year and 5 long-term monitoring projects have been initiated. Research has included studies of arthropods, birds, soil respiration, primary production, nutrient transport, geography of the “urban fringe”, a compilation of historic land use data, classification of land cover from satellite imagery, and geomorphic change in an urban river, among many others. The CAP LTER has thus far generated 28 articles (in press or submitted), over 31 presentations at national and international meetings, and 36 additional presentations.

Co-PI: Thomas H. Nash III and Corinna Gries, NSF DEB-9706984. $156,000. 06/01/97-6/00. Flora and Inventories Sonoran Desert Lichen Flora Project. The project has developed into an international collaboration of over 50 lichenologists to produce a modern, comprehensive lichen flora for the greater Sonoran Desert region (http://mgd.nacse.org/Arizona/sonoran.desert). A major emphasis of our initial funding was to increase our collections base to be representative for the whole study area. Accordingly 8 to 18 day expeditions were mounted. In addition, the project upgraded lichenological capabilities in Mexico, in part through funds from the U.S.A.I.D. program included in the initial funding. This project has supported three Ph.D. students (two completed), three post-doctoral fellows, and has hosted numerous collaborating senior researchers. The list of publications resulting from this project can be viewed at http://mgd.nacse.org/Arizona/sonoran.desert/paprson.html. A fully automated herbarium management system has been developed and is searchable at http://mgd.nacse.org/cgi-bin/qml2.0/arizona/arizonaHerb.qml. The publication of a printed flora in two volumes is scheduled for the year 2000.

P. I.: T. P. Craig, NSF BSR-9111433. $193,768. 1991-1995. Ecological and behavioral factors affecting genetic differences and gene flow in host-associated populations. We investigated how a new herbivorous insect species can evolve via a sub-population shifting to feeding on a new host plant. Our major objective was to determine if a new species can evolve in sympatry, that is without the geographic isolation of populations. Sympatric speciation is potentially a very important evolutionary process, but our studies represent one of the few empirical tests of assumptions of this model of speciation. We found that two populations of Eurosta solidaginis (Diptera: Tephritidae) that form galls on two species of goldenrod, Solidago altissima and S. gigantea, are host races. Host races are populations that are partially reproductively isolated due to their association with a host plant, and they represent an intermediate stage in sympatric speciation. Starch gel electrophoresis showed that there were significant differences in the frequencies of allozymes in the two host races indicating that they are partially reproductively isolated. The host races of Eurosta solidaginis are partially reproductively isolated because they mate on the host plant, and because they emerge at different times. Crosses between the host races produced viable and fertile F1, F2 and back-crosses, showing that gene flow is possible between the host races. Hybrids survived at a lower level than the pure host races because their unusual genotypes are not well suited to survival on either host plant. Low hybrid survival created disruptive selection for each host race to mate and oviposit on its own host plant. Our behavioral studies indicate that host plant preference is determined by a single dominant allele. This supports an important assumption of sympatric speciation theory. Our studies indicated that the host races of E. solidaginis could have originated via a host shift in sympatry. The grant resulted in the training 5 students and has produced 11 publications.