

An Internet Environment for BioDiversity Survey Collaboration and Validation

Abstract.....	1
The Information Problem.....	1
The Approach.....	1
Project sustainability.....	2
The Testing Environment.....	3
Digital Library and Portal.....	4
Sub-Collection Agent.....	4
Telecommunication and Collaboration.....	5
GIS.....	5
Education.....	6
Specimen Identification.....	8
Integrating Data and Metadata Input and Validation.....	12
Work-Centered Design and Evaluation.....	12
Knowledge Creation.....	13
User Population: Plan-IT Earth, Illinois Critical Trends Assessment Project, Illinois EcoWatch Network.....	14
Work Plan.....	15

Abstract

The proposed project addresses the problem of gathering remote data from hundreds of mostly novice collaborators in the field. This problem raises issues of distributed intelligence, field support, mobile computing, remote training, data quality and consistency, coordination of information across time and space, and real-time feedback from remote experts. We propose to build a BioDiversity Collaboration Environment (BDCE) that will build on this success and expand it through the use of information technology. BDCE addresses the need for information technology integration and information technology training for biologists.

The Information Problem

The need for new approaches to information technology development is well documented as demonstrated in a special issue of Science (Bioinformatics for Biodiversity, 2000), as well as global information projects such as Species 2000 (Bisby, 2000; Species 2000) and the Global Biodiversity Information Facility (GBIF, 2001; Edwards, Lane and Nielson, 2000). Information technology could fundamentally change the way that we understand the worlds and the way science is done. There are literally billions of specimens in museums that are underutilized. We know very little about the species on the planet, their ecology or how we are impacting them. The problem is that the right information technology does not exist, the technology is often not tailored to the task, or the information technology components are not properly integrated. We can make substantial improvements in science if the technology is carefully developed and integrated. The particular aspect of science that we wish to improve through information technology is biodiversity surveys and biodiversity education.

The Approach

In this proposal we address basic research questions about the effect of computers on collaboration and the ability of task-based software clients to match collections to the needs of particular user populations. Some of the IT that we introduce in this project will be commercial or freely available. Some will be tailored and some wholly new. It is only by this integration of old and new technology with preexisting and new work practice that we can hope to impact this branch of scientific inquiry in a meaningful way and to get an accurate picture of the effectiveness of information technology. Among new technologies discussed later are analogy-based species identification, collaborative information retrieval, proactive data harvesting and formatting for field application, use-based information repair, and extensions of data interchange standards such as ZBIG (Species Analyst, 2001).

The object of the work is to reform the way the information retrieval is done and to integrate information retrieval with information creation and organization. In the old static model of retrieval an information consumer sits in front of a terminal and passively receives information. New information technology allows for a much more dynamic model of inter-person collaboration and information capture and creation as part of an information intensive task. People retrieve information to help them perform some task. Frequently the task is complex enough and the information so sparse or ill-formed that the retrieved information does not help task performance. When this happens in the old model the person steps out of the computer information system and turns to other resources such as books and papers and most important, other people. If all goes well they will eventually accomplish their task, but the computer information system did not help with the process and was not itself improved. There is a great need for better integration of innovative IT into work practices. We know from projects such as the Alexandria Digital Library GeoLibrary that simple introduction of information technology in a single design and delivery phase will not facilitate work in all cases (Hill et al., 1997; 2000).

We propose to conduct this research in a mobile computing environment, literally in the field. This introduces interesting variables that are not normally considered in information management research. These variables include the need for extensive data caching to minimize the effects of communication bottlenecks such as digital cell modem connections. The application requires a nearly hands-free information retrieval interface. In this fieldwork, workers are usually limited as to the information resources that they can bring. We propose to use mobile computing and network technology to bring much more information to the field, potentially changing the nature of the task. We will also challenge the one user - one database model. Our information system will deliver information from multiple sources, both human and database, over limited bandwidth and relatively modest compute power in the field. Finally, the format of scientific data must make it accessible to the average users in the field.

We propose to build a system that integrates collaboration, collection development, information retrieval, and information creation and management. A key component of this technology is the development of software agents that autonomously analyze the needs of biological survey members, survey site characteristics, prior surveys of the same site, natural history museum records and other resources to create a customized collection of digital support material that can be cached on disk and taken to the field. This will help with bandwidth limitations in the field. We accomplish expanded collaboration by building interpersonal communication directly into the retrieval interface. Consumer/searchers can easily redirect information needs and search specification to remote domain experts. This may include text, images and sound. The expert may answer the immediate information need from any distance and in so doing incorporate information provided by the 'consumer' plus the solution into the database system. Information consumers then become information generators, tailoring system content to the task so that later "consumers" will not have the same problem. This IT integration can cause a qualitative shift in work practice.

The research question is whether this process improves task performance over time without introducing so much overhead cost for the participants as to negate the advantage of the innovations. To accomplish this we need to track and measure task performance, intra-system and extra-system information use and communication. Task performance can be measured by time to completion, number of tasks completed, error types, error rate, use compliance, satisfaction, and continued participation.

Project sustainability

This project will be developed with an ongoing project being conducted by the Illinois Department of Natural Resources. Any parts of the project that prove to improve the quality and volume of data without substantially increasing costs will be considered for inclusion in the main project. This means that these tools will be made available for use in hundreds of biodiversity surveys across Illinois. Other states turn to Illinois for examples of successful volunteer programs. A goal of this project would be to export the technology for use throughout the country.

The Testing Environment

An ideal test case for this problem is biodiversity management. Natural and human forces are causing rapid and extreme changes in ecosystems around the world. In order to make more astute management decisions, policy makers need systematic and objective evaluations of the state of these environments. The task is too extensive to be accomplished by professional biologists alone. Thus, there is a great need to increase the numbers of well-informed volunteers (Citizen Scientists) participating in data collection and to increase the quantity and quality of data that they produce. The Illinois Department of Natural Resources and the Plan-IT Earth project have a successful program for training volunteers and collecting data. Participants include hundreds of students in high school classrooms, as well as adult groups. We propose to build a BioDiversity Collaboration Environment that will build on this success and expand it through the use of information technology.

A major issue concerns how to structure the environment for collaboration among field researchers and central facilities. This project would develop a collaboration space in which users can represent data in multi-modal forms such as digital (micro)photography, audiograms, sketches, and text. System tools would help users perform specimen identifications and record data on their observations. These records would move through a multi-level review process in which meta-data is constructed to indicate the credibility of the data for making inferences about biodiversity.

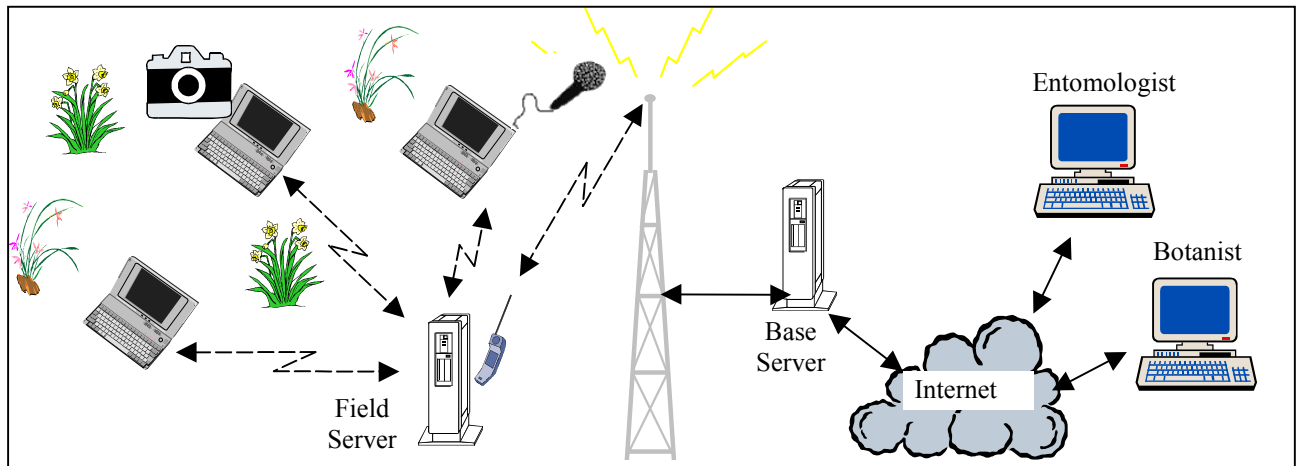


Figure 1: Communication Architecture

Specimen identification is one of the main tasks in biodiversity surveys. Unfortunately, the identification tasks performed in these surveys are costly and time consuming. The non-professionals and student professionals who are being called upon to aid in this work often find the identification process difficult and tedious. This project will use interactive keys such as LucID to aid in the process. This project will borrow heavily from the information retrieval and visualization research to explore innovative mechanisms to solve this problem. This project will provide new lines of communication between biologists and the Citizen Scientists who conduct the surveys.

The vision is that a modestly trained volunteer will be able to observe specimens in the field and enter identification, location and other data into a mobile, networked computer. This data will be instantaneously integrated with other information from the study area as well as other study areas. The volunteer could use the computer-aided instruction components of the system to learn identification characteristics of species involved in the survey. More importantly, when the volunteer runs into difficulty with specimen identification. For example, they will be able to tap into a broad range of resources that will aid in the biodiversity survey tasks. For example, they would be able to use interactive keys to attempt identification. If this fails they may use visualization environment to perform an analogy-based scan of government taxonomic databases. The volunteer may pull up digital images of prior surveys of the same location with species found there or images from other similar locations from the same time. The

volunteer may choose to load digital images or sounds into a collaborative environment that allows the volunteer to communicate with other volunteers as well as entomologists, botanists, and other biota experts around the world. Once a problem specimen has been identified, the volunteer may choose to become an author in the system. The volunteer may assemble multimedia species description information and submit it to an editorial review process. Approved materials would be added to the data collection and made available to other biodiversity survey volunteers. Volunteers would also be able to use the system as a gateway to educational information about biology, ecology, and ecosystems. Some of this information would be native to the BDCE while other information would be available through link to other Internet resources.

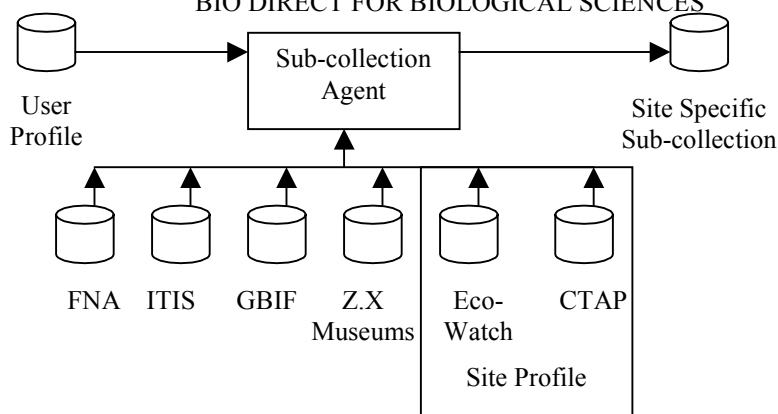
Digital Library and Portal

The collaboration environment will contain a wealth of information that may be viewed as a kind of public library but one with dynamic and animate content. Information in the library collection will not be static, but rather will include access, manipulation, and visualization tools to support specific tasks. It is these mechanisms that will attract patrons and inspire volunteers to add new content. All content will be accessible through standard World Wide Web interfaces. The central project staff will provide initial seed programs and data. The library initially will contain species description records developed for this project by the Illinois Natural History Survey. The records include descriptive information, drawings, photographs and distribution maps. The initial species represented would be those that are identified as bioindicators by the ForestWatch scientists. The collection will also include data or gateways to data from national data collections. Sample botanical collections would include the Flora of North America North of Mexico (FNA, 1999), CalFlora (1999) and Plants (USDA, NRCS, 1999), the Integrated Taxonomic Information System (ITIS), Biota of North America Program (BONAP), and ZBIG (Species Analyst, 2001), as well as others. Collection development would be guided by the species that the surveys focus on. An example graphical user interface into the Flora of North America may be found at <http://www.biobrowser.org>.

A key feature of the library is that it will be based on an open-document architecture and data inter-change standards such as Z39.50. The Z39.50 standard specifies a client/server-based protocol for Information Retrieval. This will allow us to collect data from consortia of museums using X.Z (<http://habanero.nhm.ukans.edu/TSA/start01.asp>) Also, as part of the NSF, National Science Digital Library project (<http://cecssrv1.rnet.missouri.edu/NSDLProject/>) in which the PI participates, Missouri Botanical Garden will be providing standardized access to the TROPICOS collection. This functionality allows searches for species across many museums. Where appropriate we will use the technology developed for Emerge (<http://emerge.ncsa.uiuc.edu/>) at NCSA. This means that researchers and educators will be able to write their own applications to access the data for example through NSDL. All new software for the project would be developed under the OpenSource license and the source code published. Content for the collection that is not already covered by other license agreements would be incorporated under an OpenContent license. A copy of the OpenContent license agreement may be found at <http://www.opencontent.org/home.shtml>. In short, participants and volunteers contributing information to the project will agree that the information will be freely available for anyone else to use without charge.

Sub-Collection Agent

The collection agent is autonomous software that develops a profile of the people doing a survey, a profile of the site to be surveyed and profile of the information resources available to facilitate the survey. The available data is too extensive to keep on disks in the field and bandwidth is too narrow to call much data on demand. A more intelligent approach is needed. The agent would take action to collect resources from relevant data collections. Information about the people will help define the type of materials that might be appropriate. Information about the survey site would limit the information that was needed. For example, the elevation, latitude and longitude and general habitat will help determine what flora and fauna might be encountered. In our test environment EcoWatch, the agent can extract information about the survey site from EcoWatch records. In the BIBE project we are parsing Flora of



North America data to make information such as elevation and distribution easy to extract by adding XML markup. We will need to add additional functionality in this project to extend that capability. With all of this information, combined with prior knowledge of the type content of the data collections, the agent can create a local database. It could use web-spiders from the BIBE project to extract and format species records from the Flora of North America, Butterflies of North America, USDA Plants, museums using Z.X servers, as well as the project collection created by other members of the project conducting previous surveys. This new collection would include text descriptions, maps, drawings and photographs.

Telecommunication and Collaboration

The most important resource that can be provided to people conducting biodiversity surveys is other people. The BioDiversity Collaboration Environment (BDCE) will provide for both synchronous and asynchronous communication. Scientists, biology teachers, high school students and other volunteers will all be able to communicate. Each field unit will provide a field server as a gateway to email, shared whiteboard, shared data-area for images, text and other media. The administration and staff at Prairienet have extensive experience at installing hundreds of machines into civic and learning communities (www.prairienet.org).

We will also use software and expertise that has been developed through the LEEP3 distance learning program at the Graduate School of Library and Information Science at the University of Illinois (<http://leep.lis.uiuc.edu>). Currently over 150 students participate in this program. Students from as far away as Alaska, Oregon, Massachusetts, the island of St. Croix, and Japan learn together. The technology runs on standard personal computers over standard telephone lines making it compatible with technology available in most high schools. Standard commercial products such as NetMeeting are not capable of meeting the mobile system requirements. We will use this environment as a starting point to develop Java and PHP applications to provide electronic bulletin boards, live-online Internet relay channels (IRC), as well as live and recorded audio broadcasting through RealAudio. Recorded RealAudio will be used in the mobile units of the project. Live RealAudio broadcasting would be very useful for remote education of volunteers, but we will seek funding for this aspect of the project at a later date. The advantages and disadvantages of this technology are discussed by Ruhleder (2000; Ruhleder and Twidale, 2000).

GIS

Geographic information technologies as a decision-support mechanism have been especially effective in planning information support and developing sociotechnical systems for intelligent facilitation in the context of “collective design—social interaction, interpersonal communication, and community debate that attempts to achieve collective goals and deal with collective concerns” (Klosterman, 1997, pp. 51). This role of geographic information technologies (Remote Sensing, Global Positioning Systems, GIS-based Database Management Systems, etc.) is particularly significant in its application towards building of collaboratories and exchange-networks in BDCEs. This is conceptualized in terms of geographic information technologies contributing towards an integration of specific task activities in BDCE contexts, such as data collection, information representation and visualization, retrieval, and establishing inter-person communication and collaboration environments, within geo-referenced settings. This essentially

implies that when combined with collaboration work environments and virtual community settings (both synchronous and asynchronous), geographic information technologies will be able to analyze attribute biological data collected by citizen scientists in terms of its geographic locations and the information linked to those locations. Such information will be created and visually represented within BDCE to potentially build upon the existing the database and become immediately available to other citizen scientists. This would provide manifold potential benefits, some of which are the following:

- Building geographic information technologies for BDCEs will expand the potency for representation and analysis of the biological data collected. It would provide a means to volunteers and citizen scientists for visualizing the information in terms of its location and the location of the volunteers.
- Combining this with other available online sources such as county census data, population and forest distribution profiles, land use and land cover information, GIS datasets for climate, soils, population densities, etc. will contribute towards a deeper grasp of the complexities and multidimensional environmental and ecological issues.
- Developing geographic information technologies for BDCEs will be especially significant for policy implementation, planning, decision-making, and effective organization and management of resources owing to their abilities to perform statistical analysis and modeling, as well as integrating visualization, multimedia and the Internet for communication support and information exchange.
- A geographic information technology based BDCE is also conceptualized in terms of providing an immediate visual information system for return-feedback to volunteers collecting the data. Interviews with volunteers from prior EcoWatch programs have pointed out the need for some kind of return-feedback about the data collected from the concerned agencies.

Education (Conducted as sub-contract by Eastern Illinois University)

The need to build and maintain a scientifically literate citizenry has been recognized as a national and local concern. Among the findings and recommendations of the National Science Education Standards (1996) is that youth should be familiar with the world and should recognize its diversity and unity. One very effective way to learn about the natural world is to be directly immersed in it. The outdoor environment has the potential to serve as a prolific laboratory with diverse opportunities for students to study various science-related phenomena and processes.

A greater effort to use the natural environment as a learning resource is vitally important today because of the nature of societal change. Urbanization and industrialization have created concrete enclaves that have limited proximity to open spaces and the physical environments. Because of a lack of contact with the physical world that supports them, many youngsters are now removed from natural elements and so they have little experience, knowledge, concern, or appreciation of the natural environment (Gambro and Switzky, 1996).

Similarly, formal education has primarily been conducted within the parameters of four walls. The National Science Education Standards acknowledges that "the classroom is a limited environment" and efforts should be taken "to extend the science program beyond those confines." The Standards continue to state that the physical environment in and around the school is a valuable resource and can be used as a living laboratory. Yet, teaching directly in the natural environment is a strategy that only a relatively limited number of educators employ.

In a similar context, the new standards for science education proposed by the National Research Council place increased emphasis on student-directed collaborative projects. The focus on projects is also documented in the science document, Benchmarks for Science Literacy (1993), in which student investigations are recognized as an essential part of the total science experience. Investigations are credited as a valuable aid for students to learn how science works. To meet these current expectations, teachers need experience in facilitating investigative projects that mirror scientific processes, research strategies and employ the current and relevant technologies. To address these needs for the inclusion of field-based experiences and provisions for "real science" projects and investigations, PLAN-IT EARTH

was developed by a team of science teachers and science educators in collaboration with science agency personnel. This project has provided training to over 400 science teachers and has reached over 10,000 students in the six years that it has been disseminated. This project was designed primarily to assist in the development and strengthening of teachers. Subsequently, their students' interest in science will improve as well as their background knowledge of basic science concepts, scientific methodology, issues analyzed and investigative processes through environmental, field-based learning experiences.

The primary educational goals of the project are:

- To provide teachers and students with an interdisciplinary set of inquiry-based investigations that focus on Illinois ecology;
- To develop an involved group of Illinois citizens who are interested in the environmental health and assessment of the state's and their own local ecosystems, beginning in their high-school years and continuing through life;
- To regularly monitor the changes in these ecosystems on a long-term basis, contributing to the Governor's Report on the State of the State's Ecosystems.

One facet of PLAN-IT EARTH focused on monitoring that was aligned with the efforts of the Illinois Department of Natural Resources' Critical Trends Assessment Project (CTAP). Financial support for the program primarily came from the National Science Foundation, the Illinois State Board of Education and the Illinois Board of Higher Education. Through these funded programs, teachers received the training, assistance and support for implementation of the program in their classrooms. During the past six years, the teachers demonstrated great proficiency in utilizing the classroom and field-based exercises; however, they consistently reported that they were experiencing challenges with the state monitoring procedures because of the difficulties associated with accurate identification of the targeted species. Even though the prescribed monitoring protocols were only one part of the program, the participants wanted their data to be reliable and valid indicators of the condition of the monitored site. Adequate "expert" on-site verification of the identifications was a need cited by a majority of the participants. In addition, teachers expressed their desire to extend and expand their familiarity, comfort and competence in utilizing current and appropriate technologies in environmental monitoring.

In response to these stated needs, a pilot project is proposed in which a team of five teachers from central Illinois and their students will be involved as one of the populations that will pilot the use of a digital library in the monitoring of their established forest plots. These teachers have been involved in the project and field monitoring for the past four years and have adapted the project to meet the needs of a diversity of students. These teachers have made it possible for a wide range of students to be involved in the monitoring exercises. Students with learning disabilities (Illinois School of the Deaf) as well as gifted and talented students from honors biology and chemistry (Jacksonville High School) have participated in PLAN-IT activities and monitoring exercises. Initially the cadre of teachers would receive training during the first year of the project and would work collaboratively with the field professionals as well as the students in generating and using the biodiversity information. The data would serve as a base for the teachers and their students to pursue additional investigations and research projects that could be shared through the network of PLAN-IT participants throughout the state. It is anticipated that the strategies, procedures and information resulting from the training and fieldwork would be disseminated and offered to additional clusters of school-based user groups.

As a result of this program, it is anticipated that participating :

- teachers will become skilled and knowledgeable in identifying indicator species through use of a digital library.
- teachers will demonstrate proficiency in facilitating student's learning and skill in the retrieval and analysis of data related to environmental monitoring.
- teachers and students will disseminate information, questions and findings to other populations such as scientists, other teachers and student groups.

- teachers and students will generate, conduct and communicate researchable projects related to monitoring exercises and data.

Specimen Identification

One goal of the proposed research is to produce more and better data about the distribution of species. For the people conducting the biodiversity surveys, one of the key bottlenecks to data collection is species identification. According to internal sources in the ForestWatch program, identification accuracy may be 80%. Preliminary independent field verification confirms this figure (Heidorn, 2001). The number of species that volunteers are asked to survey is kept to several dozen because of this accuracy barrier. Therefore, significant effort will be dedicated to the design, development and testing of a system to aid in identification.

While the tasks involved in conducting nature inventories is fairly well understood, the current project will introduce a number of new unique components. Most significantly, Citizen Scientists will conduct the inventories. Special attention will need to be given to the support infrastructure since the people doing the inventories are not as highly trained, as is usually the case. There are three components to the support infrastructure; training, mechanical (and computational) aids, and professional verification of data. The important characteristics of species are sometimes subtle. Research questions will include What are novices' cognitive representation of biological data such as species identification? How does this representation differ from that of experts? What vocabulary and visual aids can be used to facilitate the inventory tasks. How effective are different "high-tech" and "low-tech" techniques for the different user populations? Effectiveness can be measured in terms of accuracy and quantity of data, learning and satisfaction of the participants, and continued participation in the project. These questions are particularly important since volunteer-based data gathering can be inexpensive. To include the broadest segment of the population while ensuring the quality of the data, inventory aids are essential.

An inventory involves four main tasks: establishing a survey site, identification of a species, recording intrinsic and extrinsic facts about the specimen, and verification. In many cases the identification of the species may be the most complex of these tasks. Aids must be constructed so that Citizen Scientists can make a specimen identification with relative accuracy in a time efficient manner. The aids must also provide a means for experts to verify the data.

Intrinsic information is information about the properties of a specimen. The properties of interest will vary according to the species. The properties might include height and circumference for a tree or weight for a toad. Extrinsic information is generally about the local and regional environment. It might include the exact location within a survey grid, soil type, slope or moisture. The problem for the Citizen Scientists is to decide what data are needed. The inventory aids, independent of whether they are paper or electronic, should guide the worker through this decision.

Methods for Specimen Identification

Pankhurst (1993) describes five methods for identifying a specimen. All of these methods will be supported in the current system. For some of the methods, such as keys, we will develop and evaluate multiple approaches. As in the other studies in this project, we will collect data for different user populations and will evaluate the time requirements, accuracy, user satisfaction and costs of the techniques.

1) Know what it is already. This is certainly the quickest technique and one area where experts will differ from novices. At least for the most prevalent species training will be a valuable resource for novices and is already included in the ForestWatch program. Image intensive web-based training will be included in the collaborative environment. Still, the novices will know many fewer species and they will need to rely more heavily on other, slower techniques.

The research questions are, what is the rate of "just know" classification in different user populations using the BDCE? What are the differences between BDCE participants in ForestWatch and

ForestWatch participants not in using the BDCE? The variables include time to identification of an unknown, and error rate.

2) Ask someone else who knows. Certainly if there is an expert on hand many species may be identified in this manner. Students may use the knowledge of their own biology teachers by asking them in person. Because of the availability of telecommunications however, we are not limited by the temporal or spatial location of that knowledge. It is also possible to have the system accumulate knowledge frequently asked question documents and other sources. It is also possible to exploit prior data. On an initial inventory the probability of encountering a particular species is dependent only on the global distribution of the species. However, later inventories and individual identification tasks can be informed by prior knowledge. The inventory aids must bring this prior knowledge to the field and make it available to the Citizen Scientists there. This can be done in print or electronic format for each study. There are advantages to electronic form (such as searching) which need to be weighed against expense and tolerance of conditions in the field. The identification may also be deferred to an expert not in the field either by sending a specimen to the expert or by sending a representation such as a photograph of the unknown species. Representations may be sent electronically and potentially in real time, but this may not be cost effective.

The research questions, What is rate of use of prior inventory data and rate of deferred identification? Who do the Citizen Scientists consult? How often do they consult them? What information is exchanged? Why do they consult some sources and not others?

3) Comparison methods. Compare a specimen against a collection of named specimens. This can be done at a herbarium or by using guidebook of photographs or drawings of known specimens. As pointed out by Pankhurst, this technique is not only slow, but images are often difficult to use for numerous nondescript species such as grasses. The biota employed for ForestWatch are chosen to avoid this issue, but since the goal of this research is to expand coverage of the surveys, improving the ability to distinguish between similar species, the issue must be addressed. Electronic media do not have the same limitations as printed guidebooks. In this study we will compare the effectiveness of guidebook aids to electronic equivalents for image browsing as discussed below. We will provide drawings and photos from the data collections and from prior surveys. Unfortunately the natural history museum Z.X standard does not include provisions for exchange of diagnostic images. This is the topic of a separate IMLS proposal.

4) Use a dichotomous (diagnostic) key. This is a very popular technique. There are, however, several drawbacks. A mistake in any decision will lead to a faulty classification. The tools are even more problematic for novices since they require the knowledge of specialized vocabulary. The example that Pankhurst (1993) uses is "Sterile rosettes present". Novices are unlikely to know what this means and consequently will not be able to use the key.

In this study, we propose to test three key-like aids to determine which is best for each user population. The first is a standard dichotomous key. The second is a photographic key built directly from the verbal form of the key where phrases like "sterile rosettes" are augmented with visual examples. This is supported in the DELTA (DEscription Language for TAXonomy) character encoding standard format and IntKey (Lawrence et al., 1999). The third key-based technique that we will evaluate in this project is multi-entry keys. This is similar in format to the PANKEY system (PANKEY, 1999) and LucID (1999). PANKEY consists of different programs, which facilitate the creation of identification keys, identification of unknown specimens, production of computer-generated descriptions of plants, and searching for diagnostic characters. Multi-entry point keys such as PANKEY allow characteristics to be entered in any order. The system prompts the user for characteristics that would be useful for identification. As part of a prior NSF project (Heidorn, 2000), we are developing a web interface to Delta files by adapting PollyClave (<http://prod.library.utoronto.ca/polyclave/>) for Butterflies of Illinois.

5) Computer-based methods. These are methods that often do not have correlates with manual techniques. The non-key based method of this class that will be developed for this project is "identification by analogy".

Identification by Analogy

The most innovative computer-based method of identification developed in this project will be "Identification by Analogy." When novices and experts are asked to describe plants in natural language, their preferred descriptions are not character states (Heidorn, 1999). Instead, if they do not know the name of a specimen they will name a species that it is like, or they will identify parts of the specimen that are similar to a known species ("The leaves are like a geranium."). Some speakers, particularly novices, perhaps because they do not have a large vocabulary of plant names to pick from, will create analogies outside of the domain. So, a speaker might say, "The flower looks like a donkey's head with the ears hanging down." Experts use group characteristics to the same effect, sometimes identifying the likely genus of a plant. The advantage of these methods is that the speaker can be sure that the listener knows the terms. Another, possibly more important advantage is that the speaker is communicating the value of many lower lever characteristics all at once through the analogy. The speaker then can make minor corrections to the analogy and leave the other characteristics intact ("The flower is like a daisy, but the petals curve down").

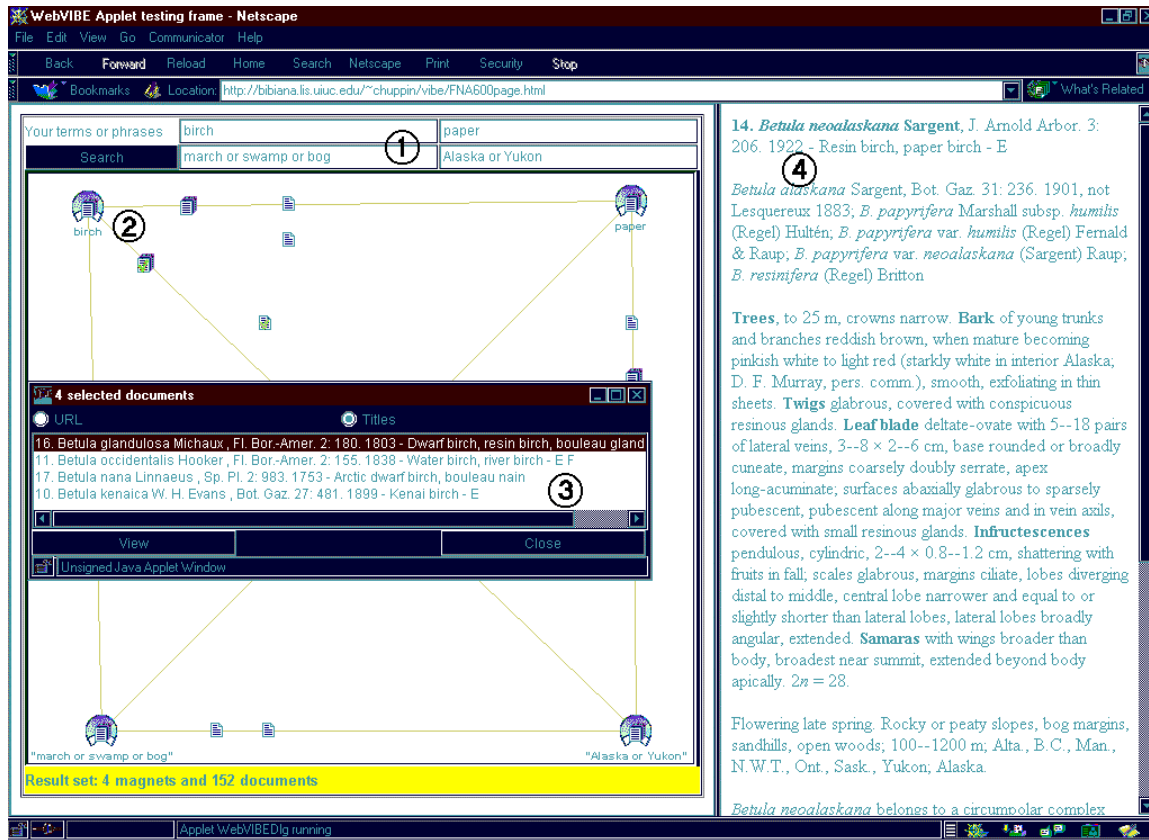


Figure 2: Biological Information Browsing Environment

In a current project (Heidorn, 2000), we are adapting a popular mechanism from information retrieval research, called user feedback and incorporate it into a graphical user interface environment. In user feedback model a user performs an information query using descriptive terms. The system returns a set of hits or items matching the search. If the proper solution is not immediately obvious the searcher marks some of the returned items as near matches and others as poor matches. The retrieval system then

combines information about these matches and the original query to form a new query for the user and then produces a new retrieval set. The work will be based in part on WebVIBE (Olson, et al., 1993; Morse & Lewis, 1997; Morse, et al., 1998).

The biological information browsing environment (BIBE) depicted in Figure 2 is a modified version of WebVIBE. A person may enter descriptive information for a species in the area marked with the Number 1 in this image. Each component of the description may be an arbitrary Boolean expression. After the user clicks the "search" button, the system displays magnet icons representing each of the search components, such as the icon to the upper-left of the Number 2 in Figure 2. Matching records describing species are represented as the rectangular icons such as that to the lower right of the Number 2. Clicking on a single icon will cause the associated species record to be displayed as on the right of the figure under the Number 4. Clicking on a stack of document icons will open a selection list displaying the titles of the descriptions, as marked with the Number 3 in the figure.

The following text provides an overview of the planned functionality for BIBE in a high speed network. The features faceted retrieval, faceted exemplar-based retrieval, and knowledge-based data extraction are being funded under a prior grant (Heidorn, 2000) Details of each of these features may be found elsewhere (Heidorn 2001). This will be followed by a discussion of features that will need to be added for the Collaboration environment.

Faceted XML retrieval: Documents in collection such as FNA are composed of discrete parts. Scientific journal articles are composed of a bibliographic section (author, title, and publication), abstract, introduction, methods, result and others. The adoption of XML standards will increase the prevalence of databases with this type of faceted structure. We expect task performance to improve if users can use these facets in query formulation and browsing. In this ITR project we propose to extend the functionality and generalizability of XML and other markup standards. Field level indexing in the BIBE project is tailored to the collections used in that project including FNA and CalFlora. In this ITR project we would adopt any available XML-QL: A Query Language for XML (Dennenberg, 1998; W3C Working Draft 31). In this ITR project we will provide forms for easy generation of XML compliant biological document to be reincorporated into our main dataset as part of the information "consumer" publishing process.

Faceted exemplar-based retrieval: In the BIBE project we are developing exemplar based retrieval functions. People can conceptually organize large collections of objects using a finite set of object prototypes or exemplars (Rosch & Mervis, 1975, Rosch et al., 1976). Exemplars are individual objects in the world that by virtue of their similarity to a group of objects may serve to represent the entire group. Prototypes, exemplars and modification functions serve as the basis for communication (Heidorn 1997, 1998a, 1998b). When a speaker describes a plant that is unknown to a listener, people frequently refer to a plant that is known to the listener and then describe the difference between the known and the unknown (Heidorn, 1999). A person may describe a plant as being "like a daisy". By pointing, a user could create a search description where "Leaf=*Quercus tardifolia*" and "Leaf=*Quercus agrifolia*". This type of relation will be supported by allowing people to use facets of document descriptions as POIs. In this example, the terms in the leaf facet of the exemplars of *Quercus* would be used to define the similarity of the POI to the other documents in the retrieval set. As is the case with full text document retrieval research, similarity will be defined by term frequency statistics. This approach has not previously been used in conjunction with biological databases and the identification task. In the mobile environment proposed here, we will need to modify this process, to cache exemplar images in the clients and the field base station to avoid communication bottleneck to the home base station and the Internet.

Four new collaborative information retrieval functions will be investigated in this research. These include **Share query**, **Share search result sets**, **Share documents** and **Submit new document**. The first three functions allow people on multiple distributed locations to share and modify their query and retrieval environment. Using this functionality, a searcher who is in doubt could ask another user for his/her opinion about a particular specimen or field location. They could ask questions like, "Do you think it is this one?" where the reference for "this one" is clear from the shared display. The collaborator might be

able to say something like, "I think it is too small. Let's add a length restriction to the query and see what happens." The collaborator could make the modification and both parties would see the result. The collaborators could be novices, experts, or a combination. The query would be shared in the extended BIBE interface. The talking and live photos of the specimen in question would occur through the IRC in the modified LEEP and commercial tools (such as firetalk).

Integrating Data and Metadata Input and Validation

There are two main types of data that may be entered into BDCE by the citizen-scientist volunteers. They are distribution data and species description support metadata. The task in this case is the collection of biodiversity data. Additional (meta)data facilitates this main task. Each data type requires data validation before it enters the main data collection. Biodiversity data will be entered through forms that are analogous to those used in Illinois's EcoWatch program; checks can be added to data fields to question volunteers about irregular entries. For verification, a certain percentage of samples from surveys are currently mailed to biologists at the Illinois Natural History Survey INHS at Urbana-Champaign campus of the University of Illinois. In the new system digital image telecommunication would be used to augment and streamline this process. Some specimens would still be mailed to INHS but online visual verification will provide much more timely feedback to volunteers, preventing error propagation. All data would initially be stored in the system in an unverified state. Unverified data might be accessible to others using the system in a limited context. It would only be accepted into the system as reliable after verification by the professional editorial staff.

A feature that will be unique to this system will be the ability of volunteers to augment species description data. For example, some digital images from the data verification process will be linked to the species description records that are used for training and identification. Both positive and negative examples could be provided to assist future volunteers with identification. Scientists, biology teachers and other volunteers would be encouraged to submit additional data that would aid in species identification and understanding. The identity of the author of the data would be made public or private at the author's discretion. In all cases, data would only be added after verification by editorial staff. The editorial staff would be motivated to make additions since the data will improve data quality through its training value. Over the long run this up-front time expenditure by editorial staff will have work for EcoWatch professionals since volunteer questions tend to center around a small number of exotic species that are not being surveyed in this project. Once data on these species is added to the system it should limit the number of questions and make answering questions easier since it will be easier to refer to the new "published" material.

Work-Centered Design and Evaluation

BDCE development will be informed by assessment and evaluation studies conducted by our social science team. Information systems design has traditionally included attention to identifying the information needs and tasks as well as the information-seeking behavior of intended system users (Lancaster, 1995; Paisley, 1968). Such studies have examined the specific needs of a particular user population, such as R&D workers (Allen, 1984), oceanographers (Hesse et al., 1993), or fiction readers (Pejtersen, 1992). A number of researchers have emphasized the importance of developing richer descriptions of the work practices and social worlds of users in order to build more effective information systems (Bowker et al., 1997; Chatman, 1992; Dervin and Nilan, 1986; Lave, 1988; Suchman, 1987). This view forms the basis of recent research on the "social informatics" of digital libraries (Bishop and Star, 1996; Kling and Rosenbaum, 1998). For example, social scientists working on the NSF sponsored Digital Library Initiative (DLI) project at Berkeley argue that the digital library (DL) is a sociotechnical artifact that "must be designed and evaluated within the larger social context of the social world in which it operates. This social world incorporates the people who use the DL, the purposes for which they use it, and its intended and unintended consequences for, not just their work, but the complex web among participants, work, tools and technology" (Schiff, Van House, and Butler, 1997, p. 161).

Viewing DLs as sociotechnical systems has led social scientists working on DL initiatives to employ multiple methods of interaction with users throughout the course of system design, testing, and evaluation. It is not coincidental that the social science teams of three DLI projects--at the University of Illinois, Berkeley, and Santa Barbara--have employed some combination of interviews, observations, usability tests, surveys, and transaction logs to develop a more complete understanding of system users and use. Data are triangulated to provide not only a more detailed picture of the nature of use overall, but to explore particular phenomena of interest, such as, in the case of the Illinois DLI project, "information convergence" (Neumann, Star, and Bowker) and "document disaggregation" (Bishop, 1998).

The development of DLs--which can combine multimedia content, new tools for retrieving and manipulating information, and more open and distributed access--has prompted system designers to consider how a single system might be customized to suit the needs of a heterogeneous user population. For example, Crane (1996) considers how the Perseus multimedia DL for classics literature research might suit the needs of both school children and seasoned scholars. Developers of a forensics DL describe how their prototype will permit consultation by both pathologists and criminal investigators (Stotts et al., 1994). A heterogeneous population is assumed in the Berkeley DL for watershed planning, where users include landowners, environmentalists, farmers, scientists, engineers, and citizens. Three target user groups were identified for Alexandria (a geographical information system) in Santa Barbara's DLI project: earth scientists, information specialists, and educators. The Illinois DLI project aims to meet the needs of faculty members, graduate students, and undergraduates in Physics, Mechanical and Industrial Engineering, and Computer Science. As in these DL projects we will tailor our system to the work practices of multiple user populations.

The first step in the assessment and evaluation work for the project will be identification of user groups that represent the diverse potential audiences for a biodiversity collaboration environment so that the work practices, information needs and system use of each group can be studied. Based on discussions with staff from the Illinois Natural History Survey, EcoWatch, University of Illinois Department of Natural Resources and Environmental Sciences, we plan to target three different populations of people who typically generate and use biodiversity information: accomplished biologists, trainers (e.g., high school biology teachers, wildlife managers, park biologists, conservation managers), and novice or amateur botanists (e.g. high school students, garden club members). Accomplished biologists will be drawn from the professional botanists, entomologists and others currently involved in the ForestWatch project. Field professionals will be drawn from the high school teachers in the Champaign-Urbana area participating in the program and the third group will be their students.

Each of these groups is interested in species identification and other uses of botanical information, but individual representatives of each group will have different goals, needs, background knowledge, vocabulary, etc. The representatives of each group will have different requirements for retrieval and manipulation of bio-diversity information. At the same time, our preliminary work suggests that the collection and use of bio-diversity survey data could benefit from better cross-group interaction. Therefore, we will also be examining the flow of information between groups and how the exchange of knowledge and resources can be enhanced. Professional scientists have many ways of working and collaborating (Palmer, 1996; Palmer, 1998), and we expect that trainers and amateur scientists will have additional needs that require support, and possibly strategies that can be modeled, in the BDCE.

Knowledge Creation

Our work-centered design and evaluation will begin with establishment of a ForestWatch User Advisory Council. At least three representatives from each user group will be recruited to join the Council. They will be asked to commit to multiple rounds of workplace interviews and to participate in usability testing. They will be kept informed of system developments and their input sought on an *ad hoc* basis. We will also use the ForestWatch Editorial Committee as a sounding board for our research.

We will extend Taylor's (1991) macro approach to understanding information-use environments to include "use scenarios" of the three user groups. Collecting data through field observation, interviews, and analysis of the existing botanical information resources, we will develop scenarios that reflect the different expertise levels and work tasks of the heterogeneous user population. This socially grounded, task oriented design process will remove categorical assumptions about the needs and behavior of botanical information seekers and present closer links between specific user and system requirements. Dervin's (1992) research argues for this type of approach, presenting a theoretical and applied framework for analyzing behavior based on the situation surrounding an individual occasion of use rather than on static user demographic or occupational characteristics. According to Carroll (1994, p. 29), use scenarios "explicitly envision and document typical and significant user activities early and continuously in the development process. Such descriptions [...] support reasoning about situations of use, even before those situations are actually created."

Hill et al. (1997) describe how the articulation of "user scenarios" informed the establishment of system requirements for the Alexandria DL. User scenarios are described as rather broad descriptions of the basic modus operandi of a particular target user group, e.g.: "Earth scientists are seeking data sets, work in a high-tech environment, have very specific search criteria, and want a DL to be tightly integrated into the working environment in which they manipulate data" (p. 235). Our three user groups are meant to generate a wide range of use scenarios. We will not develop system features that assume that just because a person is an amateur botanist, for example, she will be pursuing a particular work task or will have a certain level of expertise.

In the first year we will conduct field observations, interviews, and resource analysis to develop a preliminary set of use scenarios for all three groups. The current system development phase will focus on species identification tasks by the Citizen Scientists, an established area of concern in biodiversity survey work. Our early data collection will supplement what we have already learned about problems associated with species identification in the field, while we concentrate on developing additional scenarios that encompass a variety of tasks and information problems. The scenarios will represent various means/ends combinations that people use to locate needed botanical information. For example, desired ends may include identification of a species, obtaining a list of species in my region, determining the collection site of a specimen, or learning about the natural history and habitat requirements of a species. Information used to initiate inquiries may include knowledge of either the Latin or popular name of a species, the state where a species can be found, or knowledge of the shape of a plant specimen's leaves.

Working in consultation with the User Advisory Council, we will continue to develop use scenarios for all three groups throughout the first two years. The results will be regularly disseminated to the design team through reports and meetings and then systematically integrated into the design process. Evaluation studies will begin in Year 1 after the implementation of initial species identification features. We will conduct workplace observations to document how the new tools are used and interview to record talking users experiences with the system and their views about changes in their work process. We will follow each phase of system deployment with a round of evaluation studies, and our expectation is that by Year 2 we will be able to concentrate on communication capabilities in addition to information resource features.

User Population: Plan-IT Earth, Illinois Critical Trends Assessment Project, Illinois EcoWatch Network

Critical Trends Assessment Program (CTAP) staff and Illinois Natural History Survey scientists assisted with the development of a National Science Foundation funded high school science curriculum-Pairing Learners And Nature with Innovative Technology for the Environmental Assessment of Resources Trends and Habitats (PLAN-IT EARTH). Representatives from this program would be the use base for the proposed project.

Members of the (CTAP) will evaluate BDCE. CTAP is an on-going process to evaluate the environment of the State of the Illinois. It also provides scientific support for the Ecosystems Program under Conservation 2000, a multi-year initiative of former Governor Jim Edgar to preserve and restore Illinois ecosystems. The primary goal of CTAP is to conduct statewide and regional assessments of environmental conditions. When CTAP's first statewide assessment was completed in 1994, however, scientists reported that there was insufficient data available to adequately assess ecosystem health. The proposed Internet based environment for biological collaboration could be a significant aid to this process and could be tested with three levels of participants: trainers, volunteers (teachers), and students. BDCE could assist the Critical Trends Assessment Project, the Illinois EcoWatch Network and associated scientists and citizens in the following areas:

1. Increased volunteerism: there is a pressing need for large numbers of participants in the biology assessment program. The additional feedback and live interaction provided by BDCE would encourage volunteers to continue work with the project and attract new volunteers.
2. Verify species identification: Perhaps the most critical factor in the ecosystem monitoring process is the correct identification of indicator species. The BDCE environment would allow scientists, trainers and volunteers to verify the species they have identified or to identify species that they could not identify in the field.
3. Training tool: The BDCE system could be used as a training tool for species identification. This would, in essence, act as a reciprocal digital trainer's assistant, where the trainee can review both indicator species and non-indicator species that will be encountered within the ecosystem.
4. Assessment of training: Often flaws in training techniques are not identified until after they have resulted in errors. The BDCE system would offer a more compressed time frame, from training session to identification of flaws that would allow more immediate action to be taken. If trainees are consistently mis-identifying a particular species, this information could be captured by the BDCE system and corrections could be made to the training procedures.

Work Plan

	System Development (Heidorn)	Work-Centered Analysis (Palmer)	ForestWatch Procedure Modification (Jeffords)	Curriculum Development (Lisowski)	
Port of BIBE to Field Server and Portables	LEEP Collaboration Software	Focus Group	Material Development and Digitization	Curriculum Material Development	March
					June
					September
Collaborative Retrieval Environment	Deployment 1	Deployment 1 Task Analysis	Field Support	Field Support	December
	Deployment 2	Focus Group	Material Development	Material Development	March
		Deployment 2 Task Analysis	Field Support	Field Support	June
Collaborative Retrieval Environment	Deployment 3	Focus Group	Field Support	Field Support	September
		Deployment 3 Task Analysis			Field Support