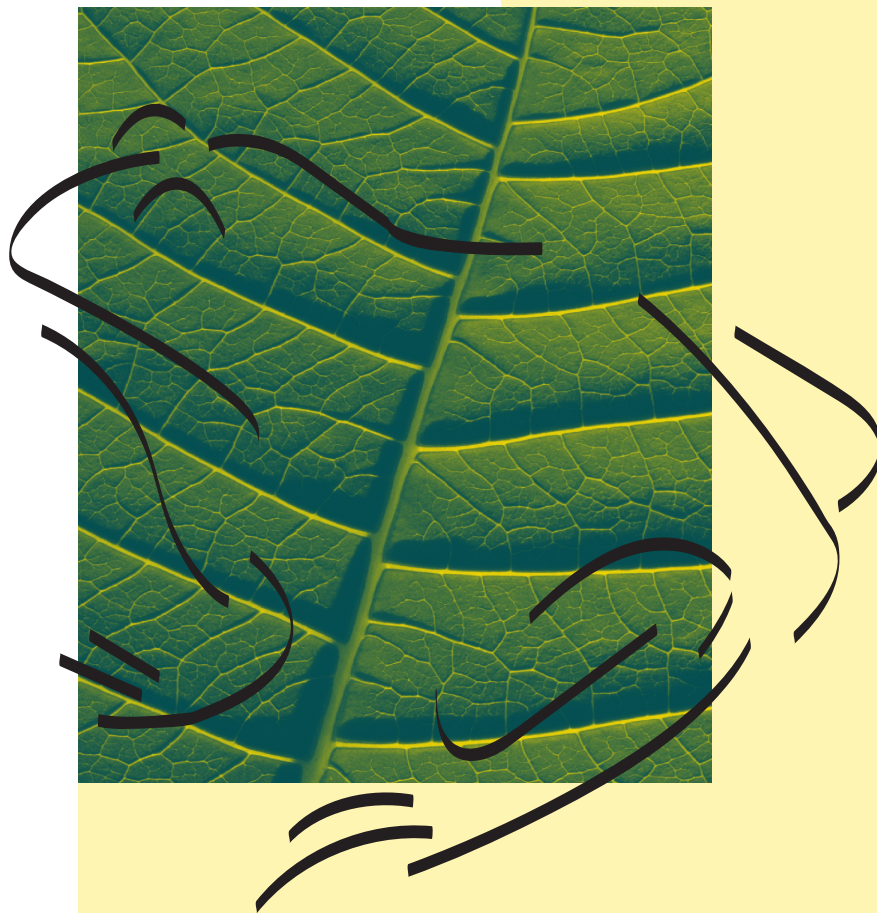


# SUMMARY OF FINDINGS



## MetaDiversity III:

Global Access For Biodiversity Through  
Integrated Systems

*By*

**Jill O'Neill  
Barbara Bauldock  
Bonnie Lawlor**

*Sponsored by the*

**U.S. Geological Survey, Biological Informatics Office  
National Biological Information Infrastructure**

*and*

**NFAIS (National Federation of Abstracting &  
Information Services)**

**March 31 – April 1, 2003  
Philadelphia, PA**



NFAIS

1518 Walnut Street, Suite 1004  
Philadelphia, PA 19102  
Phone: 215.893.1561  
Fax: 215.893.1564

E-mail: [nfais@nfais.org](mailto:nfais@nfais.org)

Web address: <<http://www.nfais.org>>

*NFAIS is a registered trademark.*



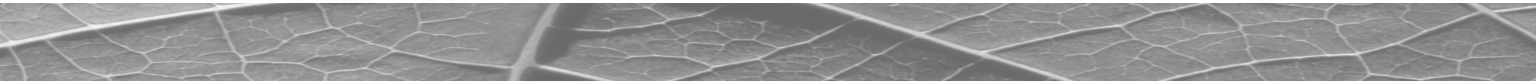
National Biological Information Infrastructure  
National Program Office  
U.S. Geological Survey  
Biological Informatics Office, MS 300  
12201 Sunrise Valley Dr  
Reston VA 20192  
Phone: 703.648.4090  
Fax: 703.648.4224

E-mail: [nbii@nbii.gov](mailto:nbii@nbii.gov)

Web address: <<http://www.nbii.gov>>

NFAIS gratefully acknowledges the support and collaboration of the U.S. Geological Survey, Biological Informatics Office, under Cooperative Agreement #02HQAG011.

*The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government.*



“All life on earth is part of one great, interdependent system. It interacts with, and depends on, the non-living components of the planet: atmosphere, oceans, freshwaters, rocks, and soils. Humanity depends totally on this community of life – this biosphere – of which we are an integral part. Biological diversity, or biodiversity, is the variety of the world’s organisms, including their genetic diversity and the assemblages they form. It is the blanket term for the natural biological wealth that undergirds human life and well-being. The breadth of the concept reflects the interrelatedness of genes, species, and ecosystems.”

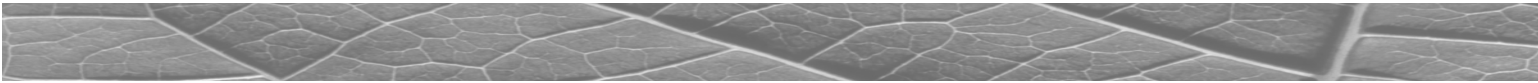
*Overview of Biodiversity, The World Resources Institute*  
<<http://www.wri.org/biodiv/bri-ntro.html>>

<b>TABLE OF CONTENTS</b>	<b>EXECUTIVE SUMMARY</b>	1
	<b>INTRODUCTION</b>	3
	MetaDiversity: Meeting the Challenge of Biodiversity Information Management	3
	MetaDiversity II: Assessing the Information Requirements of the Biodiversity Community	3
	MetaDiversity III: Global Access for Biodiversity Through Integrated Systems	4
	<b>BIODIVERSITY AND METADIVERSITY</b>	5
	<b>MEGASCIENCE AND INTEROPERABLE SYSTEMS</b>	5
	The Envisioned Global Biodiversity Information Facility	6
	GBIF's Areas of Activity	7
	GBIF-DIGIT	7
	GBIF-ECAT	7
	GBIF-DADI	7
	Discussion Points	8
	<b>CREATING STRUCTURED DATA COLLECTIONS: THE NEED AND THE TOOLS</b>	10
	Capturing Information in the Field	11
	"Citizen Scientists"	12
	Identification Keys	13
	BIBE	13
	OPENKEY	13
	TELE <sup>N</sup> ATURE	13
	POLYCLAVE AND DISCOVERLIFE.ORG	14
	Other Data Entry Initiatives	14
	Specimen Collections and Observation Data	15
	Case Study Illustrative of Challenges	16
	Discussing the Value of Metadata	17
	Geospatial Data Collections	18
	Analytical Tools from Legacy Data	19
	Biodiversity Analysis Decision Support Tool	20
	Merging Existing Databases	21

The Role of Standards: Historical and International	22
OpenURL Standard	23
Repositories and Portals	24
Biodiversity Informatics and Education	27
<b>REMAINING ISSUES AND WORKING GROUP DISCUSSIONS</b>	28
Addressing Incentives for Sharing	29
INDIVIDUAL SCIENTISTS, RESEARCHERS, AND CONSERVATIONISTS	29
THE PROFESSIONAL LEVEL: THE CULTURE OF SCIENCE	30
INSTITUTIONAL INCENTIVES (MUSEUMS, NON-GOVERNMENT ORGANIZATIONS, AND UNIVERSITIES)	30
GOVERNMENTAL LEVEL	30
INCENTIVES FOR THE COMMERCIAL SECTOR	31
Information Discovery	31
Repatriation Issues	31
Standards Revisited	32
Who Should Set Standards?	32
Problems with Standards	33
<b>END NOTE</b>	33
<b>APPENDICES</b>	34
Appendix A: Program	34
Appendix B: List of Attendees	35
Appendix C: List of Acronyms	36
Appendix D: Index of Important Organizations, Key Words, and Links to Related information on the Web	36
Additional Reading	37

**METADIVERSITY III:**

*Global Access For Biodiversity Through Integrated Systems*



## **MetaDiversity III: Global Access for Biodiversity Through Integrated Systems — Summary of Findings**

### **EXECUTIVE SUMMARY**

*MetaDiversity III* was the third in a series of symposia jointly sponsored by the National Biological Information Infrastructure (NBII) of the U.S. Geological Survey and NFAIS.

The goal of these ambitious interactive workshops has been to identify, discuss, and resolve challenges facing the biodiversity research community in their efforts to create a global information resource – a resource essential to the development of human knowledge and the preservation of biodiversity on Planet Earth. An understanding of the breadth of biodiversity may allow a fuller appreciation of the challenges that have been identified to date.

Biodiversity is defined as the identification and study of organisms (individually, as species, and as a population), their genetic relationships as well as their communal relationships in the world's ecosystems, and the impact of human activity on those relationships as evidenced by migratory patterns, species range, and behavior. The research involves a significant investment in fieldwork and other forms of primary research. It is an interdisciplinary science in that it draws from biology, botany, biochemistry, zoology, geology, earth and environmental sciences, physical and human geography, and the atmospheric sciences. And it is a mega-science in that while it can be approached at a local, regional and even national level, for some aspects, it *must* be addressed on a global scale. The international biodiversity research community includes scientists, systematists, ecologists, policy-makers, developers, educators, curators, and other natural resource managers. The sharing of knowledge among such a diverse and widespread group requires a significant technological infrastructure that has yet to be fully designed and implemented due to the scope and complexity of the issues related to biodiversity information itself.

#### **The Scope of the Research Challenge**

Approximately 1.5 million of the earth's species have been named, but another 12 million remain unnamed. This is true with regard to both plant and animal life. Only one percent of all species have been studied beyond basic nomenclature. Researchers in biodiversity have very real concerns with regard to the discovery of as yet unknown species and the identification of endangered species before all traces of their existence on this earth disappear.

The scope of this specific issue is clearly demonstrated by the results of a six-year survey of Brazilian freshwater fish completed in 2003 that identified the existence of nearly twice as many species as had been previously estimated. Even more startling was that approximately 10-15% of those species were new. It has become increasingly apparent that the biodiversity research community is in need of a global information system that would allow them to more readily identify additional gaps in current knowledge such as that shown by the aforementioned survey.

*“Information may be accumulated in files, but it must be retrieved to be of use in decision-making”*

*–Kenneth J. Arrow, Nobel Laureate*

This global system would serve as an interactive and essential conduit between field researchers and stored information. One must realize that the need for access to biodiversity information is in underdeveloped areas where the majority of fieldwork and data collecting now takes place. Yet the vast majority of this information (75%) is held in the collections of museums and other institutions in the developed world. These collections, gathered over centuries, could provide a solid baseline of biodiversity knowledge if the related information was made widely accessible. However, not all of the collections have been adequately identified, catalogued, and preserved, and not all of the information that does exist has been properly digitized. Such archiving activities are totally dependant upon the resources of the institution holding any given collection. It could very well be that the key to solving a specific biodiversity problem may already lie in field notes and labels for an existing specimen collection, but the funding required to make that information globally accessible to researchers is beyond the reach of the institutional repository. Further complicating this problem are the new collections of knowledge being assembled, including remote sensing data, geospatial data, tissue samples, audio recordings and photo-traps. These new data formats must be accommodated along with legacy data in any data-sharing infrastructure and they must be adequately preserved. Ensuring that collections are properly digitized and building the required baseline of knowledge is a major challenge for the biodiversity research community. Yet even where digital collections *do* exist, the community has no efficient method for establishing correlations between those isolated segments of essential knowledge. The Global Biodiversity Information Infrastructure is key to unifying these information resources.

### **Creating and Enhancing Structured Data Collections**

It is difficult to imagine how the enormous amount of biodiversity data yet to be collected in the field or retrieved from existing collections will actually be harvested, stored and disseminated. Organizations, working at the local and regional levels to document the biodiversity within a specific geographical area, currently assemble information resources drawn from both literature and specimen collections. They are increasingly dependent upon volunteer “citizen scientists” to accomplish their goals – citizens who care about conservation and preservation and who volunteer for field research. The creation of new digital collection tools for use by these citizen scientists and paraprofessionals in the field is facilitating the gathering of survey data. However, these isolated local and regional data repositories must be integrated into a more global information system in order to support the efforts of scientists, researchers, and policy-makers to better understand the impact of human society on the environment. Again, the Global Biodiversity Information Infrastructure is key to unifying these information resources. But the data itself must be structured in such a way to ensure that it is accessible and usable in current and future research investigations.

Usability can be reasonably guaranteed through the application of international standards that permit interoperability of systems and through the creation and application of metadata. The uniform application of metadata to collected datasets is key to enhancing the value of scientific information. Datasets enhanced by such standardized metadata are easier to share, easier to search and easier to use. The creator or custodian of the data benefits because metadata sustains the value of the information, facilitating its continued use over time and providing a return on the investment of the resources expended in gathering the data. Users benefit because they are better able to determine if data quality, accuracy and currency is sufficient for the task that they need to accomplish. Once standardized metadata has been appropriately added to content, the next challenge is to efficiently disseminate that content over global networks and interconnected systems.

### **Informational Infrastructure**

There is a well-recognized need for a commonly agreed-upon system or information portal to be appointed as the pre-eminent repository of knowledge for the biodiversity community in order to avoid balkanization of resources. From the perspective of educators, curators and researchers, it would be useful to have a single organization or agency given the authority and on-going responsibility for the creation and maintenance of the biodiversity

information portal. There seemed to be a general agreement among the *MetaDiversity III* conference attendees that the National Biological Information Infrastructure (NBII) is the logical choice to be that designated authority and one who could represent the United States with regard to the further development of the Global Biodiversity Information Infrastructure.

### **Issues and Working Group Discussions**

The discussions that took place during the *MetaDiversity III* workshop centered around two aspects critical to the success of a global information resource. One working group focused on the creation and acceptance of international standards, a key element of ensuring interoperability between systems created in disparate settings. And it was agreed that continued involvement of the United States in the international community process of creating acceptable standards is key to the success of a global biodiversity information resource.

The other working group focused on the need for incentives that would encourage the sharing and depositing of data across institutional and international boundaries. The group agreed that the creation of such incentives requires approaches at five levels – individual, professional, commercial, institutional, and national. And they determined that avenues of approach include, but are not limited to, requiring deposits of data for purposes of accreditation and promotion, support for new forms of digital publication, and further development of a Biodiversity Information Commons to inhibit the bio-prospecting in and exploitation of under-developed and transitioning nations.

The *MetaDiversity III* symposium concluded with a unanimous agreement that the creation and immediate accessibility of well-structured text and/or multimedia biodiversity databases through a central platform would have numerous benefits for researchers and field scientists. Access to such information would facilitate the expansion of knowledge in support of better decision-making processes, would create a better quality of life for citizens worldwide, and would provide for the immediate transfer of knowledge to the global research community. Such is the goal of the Biodiversity Research Community – and such is the ultimate objective of the Metadiversity symposia that have provided a forum in which to further discussion, foster collaboration and improve the flow of biodiversity information.



## INTRODUCTION

*MetaDiversity III: Global Access for Biodiversity Through Integrated Systems* was held in Philadelphia, Pennsylvania, March 31–April 1, 2003, jointly sponsored by the U.S. Geological Survey (USGS) and NFAIS, an international membership organization for those who create, aggregate or facilitate access to information. The meeting was the third in a series of symposia focused on the creation of a fully-integrated, fully-interoperable biodiversity information system.

### **MetaDiversity: Meeting the Challenge of Biodiversity Information Management**

In the late 90's, concern arose within the research community with regard to the management of the widely diverse information related to biodiversity studies. "Biodiversity information" or information about the number and variety of living organisms, their genetic make-up, and the ecosystems in which they reside, was and continues to be an international priority. Diverse types of information and research data, including geospatial data, museum specimen collections, videos and audio files, and digital satellite images, require appropriate organization and seamless integration in order to facilitate efficient access, search and retrieval of the data. Recognizing this challenge, the USGS and NFAIS hosted the initial symposium, *MetaDiversity I*, in an attempt to bring together the many organizations and individuals who are working to create operable and interoperable information management systems that would facilitate access to this broad spectrum of digital knowledge. The name "MetaDiversity" itself referred to the use of metadata (information about information) as applied to the tasks of discovering and retrieving biodiversity information from online systems.

The discussions that took place during *MetaDiversity I*, held over a three-day period in November of 1998, focused on five critical areas that had to be addressed if the envisioned information system were to be made into a reality. Those five areas were Leadership, Technology, Standards for Biodiversity Data, Funding & Economics, and Users. Drawing from the international perspectives present at the conference, participants broke into working groups to review the specific challenges associated with each of these aspects of building an integrated information infrastructure – an infrastructure essential to support scientific advancement in this field.

The proceedings volume of *MetaDiversity I* was published both in print form and on the Web in the interests of widely disseminating the recommendations for further action to a global audience. (While the print version is no longer available, the Web-published proceedings are accessible on the NFAIS Web site at [http://www.nfaais.org/publications/metadiversity\\_preprints\\_contents.htm](http://www.nfaais.org/publications/metadiversity_preprints_contents.htm).)

Emerging from the consensus reached at the initial MetaDiversity meeting was the recognition that three action items needed to be carried through in order to ultimately meet the biodiversity information management challenge:

- A "call to community", an understanding across the full spectrum of participants (policy-makers, researchers, information professionals, educators, etc.) that such an infrastructure would only be created through consensus and a committed, cooperative effort;
- Development of standards for creation of metadata that would support system interoperability; and
- An assessment of the projected user community and user requirements for the envisioned biodiversity information system.

It was the third action item that formed the basis for *MetaDiversity II*.

### **MetaDiversity II: Assessing the Information Requirements of the Biodiversity Community**

The second MetaDiversity event was held two and a half years later in June 2001 with the planned objective of:

- Identifying the areas where biodiversity information needs were being met;
- Identifying those areas where biodiversity information needs were not being met; and
- Identifying potential user groups and the worldwide target audience for biodiversity information.

The participants in this second event, which was held in Charleston, SC, were individuals from federal agencies, academic institutions, libraries, universities, research facilities and international groups with an interest in biodiversity informatics. The diversity of the participants helped attendees to identify an equally diverse community of users made up of policy makers, ecologists, geologists, systematists, curators, foresters, land use planners, resource managers, educators (all levels), wildlife and fisheries management, consultants, funding bodies, and content providers, and others within the research community.

Attendees spent a significant portion of the *MetaDiversity II* meeting hypothesizing about an ideal knowledge environment for users of biodiversity data and systems. Eagerly, the participants assembled a “wish list,” specifying desired functionalities for a system that would allow users from this varied audience to accomplish research and field tasks, advancing scientific knowledge while furthering appropriate global development and environmental objectives.

As with *MetaDiversity I*, over the course of the two days a number of recommendations emerged from the working groups’ discussion for the further development of the USGS-led National Biological Information Infrastructure (NBII) in particular and for the biodiversity community in general. These recommendations included requests for:

- Guidelines and tools for collecting, managing and archiving data;
- Improved visualization tools, virtual laboratories, and methods of handling three-dimensional objects, such as specimens over the Web;
- Knowledge navigation tools (authority sources, thesauri, data mining and analysis software) and systems for attribution, accreditation, and rights management; and
- Support for such global biodiversity systems as the Global Biodiversity Information Facility (GBIF), the Natural Science Collections Alliance, the Global Geospatial Data Infrastructure, and similar initiatives.

All of these recommendations rely upon the validity of two assumptions: (1) That all participants across all sectors and at all levels are willing to share research data and accumulated collections with other institutions and nations; and (2) that a fully interoperable system would be developed in order to provide the technological infrastructure for that sharing of data.

In the interests of assessing the progress to be made on these two points (sharing both intellectually and technologically), the attendees of *MetaDiversity II* agreed that a third *MetaDiversity* event would be in order.

[The Summary of Findings from *MetaDiversity II* are available via the Web at <<http://www.nfais.org/publications/metadiversityII.pdf>> and at <<http://www.nbii.gov/about/pubs/metadiv2.pdf>>. The printed version may be requested from NFAIS or from NBII, while quantities last.]

### **MetaDiversity III: Global Access for Biodiversity Through Integrated Systems**

A call for papers went out in January of 2002, soliciting participation from members of the international biodiversity and information technology communities for the third *MetaDiversity* event. The scope of this meeting encompassed the challenges of aggregating, organizing, standardizing, digitizing, and disseminating all forms of content in support of a coordinated biodiversity information network.

In approaching the challenges involved with the creation of interoperable biodiversity information systems, *MetaDiversity III* opened with presentations from individuals and organizations that were already immersed in the development of information systems actively in use by the community. The initial sessions of *MetaDiversity III*, which began on Monday, March 31, 2003, were structured around global and organizational initiatives which are successfully delivering diverse forms of information and data to specific sets of users. An overview of curricula in support of biodiversity informatics was also incorporated into the morning’s program, indicating the sophisticated level attained by these information resources.

The program then turned to prominent researchers who were developing their own information tools in order to gather and deliver field data to others in the biodiversity community. Impatient with the need to push the progress of their work forward, these researchers talked about the use of mass-market software applications (such as FileMaker) to input data while allowing others to retrieve it. Questions and discussion points regarding the wisest approach towards developing standards emerged as the day went forward. (Reflecting this concern, later in the course of the event, one working group devoted themselves entirely to a discussion of this topic and offered recommendations for moving forward to resolve the concerns related to standards implementation.)

The final sessions of the program, held on Tuesday, April 1, 2003, included a discussion of the OpenURL standard and applications of metadata in a variety of settings. The final hours of the two-day event were spent by participants in two highly interactive working groups, one focused on the use of standards and a second focused on how to encourage the sharing of research findings and data by individuals and by institutions in support of global information initiatives and objectives.

This report documents the discussions held at the *MetaDiversity III* meeting and synthesizes the recommendations of the participants in the interests of furthering future developments in this field.

## **BIODIVERSITY AND METADIVERSITY**

*A researcher in New York's Central Park spots a species of caterpillar and is puzzled by its presence, as the species is not native to the region. Is this an isolated instance or has the species expanded its range?*

*A developer in Florida has plans for a new community. He is fairly confident that this is "smart growth" but wants to investigate how the addition of the new homes may impact water resources in terms of run-off. He is also looking for historical data on flooding in the region.*

*There is the potential for a spruce budworm outbreak in Minnesota. State forestry managers are alarmed and wonder if the balsam fir population has become over-mature and requires thinning out.*

The three scenarios proposed above are all legitimate possibilities. While all three examples reference the continental United States, the questions and concerns raised typify the concerns faced by developed and developing nations alike when dealing with the conservation and appropriate use of environmental resources.

Biodiversity, as both the inventory of life on the planet and as a study of the impact of human activity on the rest of the ecosystem, requires that humans capture and analyze data in order to understand what variety of life co-exists on the planet, as well as to conserve and use the variety of resources available for improving the quality of life for all.

The sustainability of our environment is an important concern for all. Cooperative sharing of data facilitates informed policy – and decision-making by the worldwide community in changing human activities that have a negative impact on plants, animals and ecosystems while supporting the economic, social, and cultural well-being of the world's populations.

Information tools and systems can now be created that provide access to data without regard to time zones or geographical barriers. It is now possible to build global systems and analytical tools that promote the use of existing data collections, linking the work of field researchers in Sri Lanka, for example, with a scientific community in the United States that, in turn, may be used by conservationists in northern Europe, Africa and Central Asia. As a result, knowledge is expanded more widely and at a far more accelerated rate.

Supporting enhanced global access to data requires that systems be interoperable and that data be flexibly structured. In the case of the former, systems must be able to support readily the exchange of data

across disparate types of hardware and software, as well as across geographic boundaries and diverse scientific communities. In the case of the flexible structuring of data, such a practice allows the re-use of data over longer periods of time and without reference to geospatial divides. The data collected and stored in one location may have value in a broad variety of applications. The standardization of those data facilitates broad re-use; links can be constructed from datasets compiled by individual researchers to other relevant content, such as taxonomic databases. Users gain when knowledge is sensibly aggregated and barriers to access are minimized.

*MetaDiversity III: Global Access to Biodiversity Through Integrated Systems* provided a forum in which experts in the field of biodiversity identified and discussed practical measures to be taken in order to facilitate enhanced access to data in a wide variety of formats through standards and through encouragement to share.

## **MEGASCIENCE AND INTEROPERABLE SYSTEMS**

Due to its very nature, biodiversity research must be approached on a global scale. The impact of a wide variety of human activities on Planet Earth's complex ecosystems and living organisms cannot be properly assessed without a view of the whole.

Researchers approach the study of biodiversity from diverse levels and viewpoints – from the genetic study of an individual organism through the Linnean hierarchy to the complexities of diverse ecosystems spanning geographic boundaries. Emerging knowledge has demonstrated the scope of biodiversity's interdisciplinary reach. It has identified problems resulting from duplicative efforts, as well as from incomplete studies and conflicting taxonomies. But neither the existing technology nor the research community has been able to readily resolve these issues. Individuals and organizations have found that they can create information readily enough, capturing data from the work of expeditions, photographs, notes, etc. But until the advent of computing power and other technological enhancements of recent decades, there has simply been no methodology for readily integrating all of the accumulated knowledge into a single information system that worldwide scientists and policy-makers might draw from in order to develop global solutions. Too high a cost, too much data, and too great a reluctance on the part of individuals, organizations, and nations to share data challenged the development of the envisioned information resource. However, there is a growing international awareness that the problems of biodiversity must be addressed on a global basis for the purposes of economic development, environmental decision-

making, and the expansion of scientific knowledge. Clearly, some form of electronic information environment must be built. The scope of the challenges pushes the project into the realm of *megascience*.

According to *MetaDiversity III* keynote speaker **Dr. Meredith Lane, Communications Officer, Global Biodiversity Information Facility (GBIF), Denmark** a megascience effort can be characterized by the following:

- The research question is interdisciplinary in scope and encompasses many fields of study;
- The research cannot be adequately addressed by the efforts of only a single country, given the expense, the expertise necessary or the scope of the data required;
- The research can be done on a regional or national level, but at some point demands global participation or components;
- The research requires involvement from both public and private sectors, including governments, academic institutions, private corporations, societies, etc;
- The research requires collaboration from many scientists and others;
- The research involves primary research as well as the development of an infrastructure; and
- The research is performed by individual scientists in physical as well as virtual settings or facilities.

### **The Envisioned Global Biodiversity Information Facility**

GBIF <[www.gbif.org](http://www.gbif.org)> certainly qualifies as a megascience effort, as biodiversity research encompasses all of the elements enumerated by Lane.

According to its mission statement, GBIF has the overriding objective of ensuring that *primary* scientific data with regard to biodiversity is openly and freely accessible to everyone, no matter where on the globe they are located. It exists to promote standards and software tools in diverse languages and in diverse character sets and computer encoding that facilitate the use of biodiversity data. Twenty-two voting participants (nations) and 32 associate participants have made a commitment to this initiative by signing its Memorandum of Understanding and by establishing a GBIF node. The nodes are part of a distributed network, as GBIF is a facility that is distributed around the planet, with its many parts connected by the Internet.

In order to promote the sharing and use of scientific biodiversity data, GBIF focuses on four areas of activity:

- Data Access and Database Interoperability (DADI);
- Digitization of Natural History Collections (DIGIT);
- Electronic Catalog of Names of Known Organisms (ECAT); and
- Outreach and Capacity Building (OCB).

The types of data that GBIF will provide are unique, as the project attempts to avoid the duplication of any existing effort. The areas for which GBIF is responsible include the registration of observation and specimen data, as well as the cataloging and search engine functionalities. However, the databases that will be included in that registry, such as GenBank or other existing databases of geospatial data, ecosystems data, climate data, and ecological data, are those generated by the global community of researchers, professionals, and other participants in the field.

Lane used GenBank as an example of how the system will operate. Until now, it has been impossible for a researcher to combine data from GenBank with specimen and ecological data from other sources without performing tedious and painstaking work. The interoperability of the GBIF information architectural components will allow data to be drawn easily from multiple sources with a single query and combined as needed by scientists and others.

The central importance of GBIF lies in its stated mission to “make *primary* scientific data about biodiversity openly and freely accessible.” In the face of increasing demands for this type of data to enhance decision making across public and private sectors, GBIF has been established to redress the inequality of data distribution distributed across the developing and developed world. By emphasizing the equitable sharing of data, GBIF will be able to facilitate the combined use of the actual environments available in which to study biodiversity patterns in underdeveloped nations with the abundance of biodiversity data already accumulated and stored in the technologically advanced developed world. Additionally, the GBIF information system, as currently envisioned, ensures that the global community will benefit from the availability of the data via a fully interoperable system of databases and other collections.

Lane further emphasized the megascience aspect of the GBIF mission by speaking about GBIF’s focus on capturing *primary* data that, because of the difficulty or time required to access, is at present not often used in natural resource policy or management decisions. Primary data that is available in digital format can significantly improve decision-making. Lane provided the example of the

National Commission for the Knowledge and Use of Biodiversity (CONABIO) that has created and maintains a national biodiversity information system for all of Mexico. The CONABIO database, consisting of more than 750,000 records of primary data, has already served as the basis for the establishment of national priorities regarding the conservation of Mexico's natural resources.

### **GBIF's Areas of Activity**

#### **GBIF-DIGIT**

The GBIF project for the Digitization of Natural History Collections (DIGIT) has a long-term vision to facilitate the expansion of biodiversity knowledge through the digitization and distribution of legacy and newly acquired primary species occurrence data. The goals of this project include the facilitation of access to data associated with the specimens in the world's natural history collections, the identification of efficient and cost effective ways to organize and accelerate the specimen digitization process, the repatriation of specimen data from the developed to the developing world, and the advancement of biodiversity science through improved access to primary species occurrence data.

It has been estimated that there are in excess of 3 billion specimens in the world's natural history collections. The DIGIT project <<http://www.gbif.org/prog/digit>> will incorporate these collections, culture collections, and observational species occurrence data as well as new data from ongoing data acquisition. With this body of information available in digital format, a quicker and more efficient analysis of the current status of biodiversity knowledge will be facilitated, resulting in a well-informed prioritization of research.

#### **GBIF-ECAT**

Closely aligned with the DIGIT project will be an effort to develop a comprehensive electronic listing of the names of all species known to science – a listing that will also accommodate various sorts of classifications – the ECAT work program.

The critical nature of this arm of the GBIF activity lies in the central importance of nomenclature in scientific literature and databases. The only common data field across a wide variety of biodiversity information databases (specimen databases, GenBank, etc.) is the field incorporating the scientific name of an organism. Without a scientific name, there is no access to the primary data. And while there are primary scientific names, the biodiversity research professional is well aware of variant names that may also be in use. Lane used the example of *Amphiachyris dracunculoides*, a species that may also be identified as *Gutierrezia dracunculoides* or as *G. texana var. dracunculoides*. Thus any system hoping to facilitate

access to biodiversity data must have incorporate a form of controlled vocabulary based upon the scientific nomenclature. It is indicative of the importance of this particular initiative that a full quarter of the GBIF budget will be spent on the ECAT project.

The long-term vision for ECAT <<http://www.gbif.org/prog/ecat>> is that it will facilitate the exploration and rapid expansion of biodiversity knowledge by providing a complete, digital listing of the names of all known organisms. One side benefit of such a digital listing is that it will be easier for the biodiversity discovery scientist to determine if a name for a particular organism has been used before. This will eliminate the problem of duplicative names for organisms, a problem that can be a significant barrier to the development of a comprehensive map of living organisms in human knowledge. It will also alleviate the homonym problem that has grown over the course of the past 250 years and permit better treatment of the synonymy question.

The participating nodes of GBIF are involved with the development of Global Species Databases, concept-driven taxonomies and listings of names that incorporate regional lists and nomenclators. ECAT will expand the electronic list of names and known organisms by working with the Catalogue of Life Consortium, consisting of the Species 2000 project <<http://www.sp2000.org>> and the Integrated Taxonomic Information System <<http://www.itis.usda.gov>>, to speed up the development and digital availability of these Global Species Databases. Emphasizing the value of regional data gathering efforts, Lane also referenced the inclusion of the regional lists and nomenclators into the ECAT. In conjunction with DIGIT and DADI, the creation of ECAT supports the total interoperability of the GBIF information architecture.

The estimated completion of ECAT, given the amount of work required to bring the Global Species Databases online and to identify authoritative lists of names and nomenclators that can also be brought on line, may be a full decade away. But, Lane assured the audience, without ECAT, the projected availability of such a catalogue would be much further out in time.

#### **GBIF-DADI**

Biological entities are far more complex than all other physical and chemical entities for several reasons. At all levels of organization, each biological entity is unique. Phylogenetic (genealogical) history matters a great deal. Biological relationships are contingent upon (and therefore complicated by) phylogeny, ecology (including symbioses and parasite relationships), chemical competencies, sensibility, vagility and mobility, etc. As a result, human knowledge of most species' natural history and systematics is highly imperfect.

Even more frustrating is that, while we know that human understanding of life on this planet does not properly encompass the reality, the current understanding and organization of human knowledge is the only available basis for decision-making, however flawed that foundation may be.

In biodiversity information, as previously stated, all biological entities are tagged with a scientific name. As Lane put it, “Phylogenetic argumentation aside, everything known about biodiversity is indexed by scientific name. That being the case, we cannot afford to dispense with the Linnean hierarchy, however flawed, too soon.” Biological understanding is complicated by intricate interactions across all levels of organization and the intricate interactions among entities within the same level of organization. This understanding is even further complicated by the inconsistent and idiosyncratic methods of data collection employed throughout history and across disciplines.

As documented in the findings of *MetaDiversity II*, users of biodiversity data and information may be involved in conservation, regulation, sustainable development, education, research or a variety of other projects for industry and the general public. The GBIF vision for biodiversity information is that the primary data accessed through GBIF will be available for all sorts of purposes. The re-use of aggregated data housed in the robust information architecture being built by GBIF will eliminate duplicate research efforts and expenditures, support interoperability between heterogeneous collections, and ultimately facilitate data-mining of the accumulated information. Users will be able to make associations and correlations that are impossible without dynamic interaction with the data. Ultimately, the gaps in biodiversity knowledge will be narrowed and possibly closed. As Lane phrased it, “If we can get a handle on what we don’t know because we have a system that is working dynamically and much more rapidly, then we will have made a significant step forward.”

Data Access and Data Interoperability (DADI) is the project that will allow data flow across the GBIF network of nodes. DADI has the objective of facilitating interoperability by:

- Working with existing data and metadata development efforts;
- Employing the most up-to-date and useful methodologies;
- Anticipating the growth of user needs/demands and available information technologies; and
- Establishing an information architecture that can evolve and adapt to changing needs and attitudes.

The DADI initiative <<http://www.gbif.org/prog/dadi>> is intended to build a system sufficiently flexible so that it does not have to be redesigned as time and technology progresses but can be simply modified and adjusted to meet evolving needs.

How can GBIF data be made interoperable? By devising and/or adopting common data and metadata standards and software protocols, and by carefully thinking through the individual logical and logistical steps needed to accomplish the larger goal. This is done via templates for “use cases.”

A variety of use case query templates have been developed by GBIF. Just one example put forth by Lane was a basic use case labeled “Find Global Specimens/Observations.” This phrase is translated as a query to retrieve specimen/observation records for a given species or location held by any part of the GBIF network. Use cases cover basic requests for retrieval of scientific or vernacular names, map ranges, etc., and can be extended to more advanced requests such as “Relate forestry practices to threatened species.” (Those seeking greater detail about use case templates can access the documentation by going to <[http://circa.gbif.net/Public/irc/gbif/dadi/library?!=/cases\\_gbif\\_network.>](http://circa.gbif.net/Public/irc/gbif/dadi/library?!=/cases_gbif_network.>)

Lane pointed out that the GBIF timeline currently in place has a prioritized set of use cases that serves as a benchmark for the GBIF network implementation as of March, 2003, and the first iteration of the full information architecture to be in place by December 2003.

### **Discussion Points**

Participants at *MetaDiversity III* had a number of questions about the GBIF vision and its goals. Ken Klemow of Wilkes University commented that data is collected with a lot of noise. Errors creep in with the collection and dissemination of data. Klemow asked what responsibility GBIF has with regard to quality control and quality assurance.

Lane responded, “Data collection is kept as low in the hierarchy as possible so that the responsibility of QA/QC lies primarily with the participant nodes. That responsibility is subsequently enforced by the nodes with their data providers.” She went on to reference the map of Mexico showing points of data collection for the CONABIO database. Speaking about the Mexican experience, Lane pointed out that the more than 700,000 records incorporated in the CONABIO database represent only about 60% of the total number of records that were gathered. “They don’t use the ones that are clearly too messy. And even with the elimination of ‘noise’ there is still a tremendously robust body of information remaining.” Another

aspect of cleaning up the data involves development of data cleansing routines specifically aimed at imprecise geographical descriptions found in specimen labels. Software is being developed, which will be available through the GBIF portal, that will retrieve and allow correction or elimination of “dirty data” or errors. Lane pointed out that in information systems used by professionals, the data will end up being cleansed by the sheer volume of use, as researchers alert data gatherers to corrections that need to be made.

*The amount of work to be done in cataloging species demands the concerted effort of entire citizen populations around the globe.*

Another audience member asked Lane whether GBIF had any programs specifically devoted towards working on the cooperative sharing of data. Lane agreed that incentives for open sharing of data and for allowance of appropriate data attribution were important issues, ones that the community as a whole must discuss. GBIF saw these discussions as part of work programs further down the road dealing with digital libraries. Said Lane, “We must work out ways for allowing someone as much credit for the creation and maintenance of a database as the current credit received for a monograph or for publishing a written paper.”

John Pickering of DiscoverLife.org pointed out to Lane that, based on the architecture of the GBIF system, the data registry seemed to be the key to everything. Pickering asked whether he, as a data provider, should register DiscoverLife.org as a single entity, or should he register DiscoverLife.org as having information about this individual species and again for this other individual species and again for a third individual species, etc. He also inquired about those instances where a data provider points to third party or remote databases for some of the information included in a resource.

Lane replied, “The registry actually has a very small number of data fields. When a query comes into the portal, it first asks which among the many data providers has information on this particular organism. The GBIF portal will check the registry and will be able to know that it only has to ping five providers with mega-databases rather than 95 separate databases, because the registry will know this from the metadata coming into the system.” She pointed out that a meta-generator is part of the toolkit being passed out to participant nodes. The participant nodes, in turn, will assist their data providers nodes in determining how such information is reported.

Questioned about the long-term financial viability of GBIF, Lane responded, “When GBIF-wide capabilities come online and people

really start using them, then the value of financial investments will be more readily seen. If the system breaks at that point because the funding isn’t there, the research communities will insist that the funding be reinstated.” GBIF also is considering expanding requests for funding to foundations and industry for in-kind contributions. Lane pointed out that “It’s the old chicken and the egg kind of thing – you need to get some money to do something, but you need to do something to get money.” When the third-year review of GBIF is issued, it should document that good things have come out of the effort. Those countries that are associate members (i.e., ones not having a financial commitment to GBIF) will then see the benefit in making a financial commitment. “We’re also looking into having a supplementary fund to which any individual will be able to contribute,” Lane added.

Bryan Heidorn of the University of Illinois, Urbana-Champaign, asked about the actual distribution of the system, specifically whether GBIF was making multiple copies for security, access, longevity, sustainability, etc.

Lane confirmed that the system was mirrored in more than one place. The GBIF registry is, in fact, a meta-registry. Presumably the nodes will be encouraged to mirror the registry so that there will be redundancy to provide backup.

Cynthia Parr of the University of Maryland asked about the relationship between the GBIF portal and other portals such as the one provided by the National Biological Information Infrastructure (NBII) at the USGS <<http://my.nbii.gov>>. As a specific example, Parr asked if DiscoverLife.org could register with the NBII and then have NBII register with GBIF.

Lane responded, “What GBIF is hoping is that providers such as DiscoverLife.org will be registering through their national node (which, in the case of the United States, is NBII).” The GBIF central registry, then, would not be dealing with a mix of registrants, some primary and some meta-registers, which might cause confusion. The Memorandum of Understanding signed by all GBIF participants says that a participant country can establish one or more nodes internally. Germany, for example, has eight. “The United States has NBII,” Lane stated, “but the US could also name an organization such as the Natural Science Collections Alliance to act as a registry specifically for museums, and other kinds of providers might go to NBII. It depends very much on how the national community wants to handle it.”

## **CREATING STRUCTURED DATA COLLECTIONS: THE NEED AND THE TOOLS**

It is an amazing fact that most of the Earth's species are unnamed. **Dr. P. Bryan Heidorn, University of Illinois, Urbana-Champaign**, offered the interesting statistics that, while approximately 1.5 million species have scientific designations, more than 12 million remain unnamed, and less than 1% are studied beyond the point of an established name. This fact is true with regard to both plant and animal life.

An interesting example of Heidorn's comment can be seen in the results of a recent 6-year, government-funded survey of Brazilian freshwater fish by researchers from four major Brazilian universities. After collecting more than 50,000 specimens, the preliminary data indicated that nearly twice as many freshwater species were living in Brazilian rivers, lakes, and streams as had been previously estimated. About 10–15% of those collected were species entirely new to science. An Associated Press story quoted the chief investigator, Dr. Neircio Menezes, University of Sao Paulo, as saying, "We have to document what exists as rapidly as possible. We are convinced there are species that have already disappeared without being scientifically described."

Even as Meredith Lane's presentation touched on the need for an information system that would allow researchers to get a "better handle on what it is we don't know," Heidorn provided a brief sampling of what it is that the biodiversity research community does understand regarding its own inadequate grasp of knowledge about life on the planet.

"Many people believe that Linnaeus completed the task of systematically organizing species. They're not aware that there is this interesting and critical problem still to be resolved," stated Heidorn. "Our taxonomic structure has fallen apart in the past 10 years, given what we know now from DNA. Our phylogenetic relationships are off."

Heidorn's statistics suggested that majority of species remain unnamed. As an example, Heidorn discussed the status of human knowledge with regard to insects. There are believed to be between 80 million and 100 million species of insects. Only 950,000 of these species have been described. New species are still being found: a purple centipede was found in Pittsburgh, PA, in June 2001, and another new centipede was found in New York City's Central Park in 2002. How can we be sure that a new species has been located? There is no comprehensive, centralized, readily available index—yet.

There are too few taxonomists to accomplish the task. The amount of work to be done in cataloging species demands the concerted

effort of entire citizen populations around the globe. "We need to capture that knowledge," said Heidorn.

One of the most prominent tasks facing scientists involved in biodiversity is that of identifying and naming species through inventories of specimen collections as well as through surveys in the field. One of the unique aspects of biodiversity research is that it is not only specimen-based (referring to collections held by museums and other institutions), but also field-based. Unfortunately, 75% of the need for biodiversity information is in underdeveloped parts of the world (where the majority of field work is taking place), while 75% of accumulated biodiversity information resides in the developed part of the world in museums, universities, government institutions, and other organizations. The creation of an interactive conduit between the field and stored information is essential.

Museums, educational institutions, laboratories, herbaria, botanical gardens, and research institutes store an immense body of cumulative knowledge. Reflecting the practices of the "old" natural history, more than 3 billion specimens are held in 6,500 natural history museums alone. However, the captured data and information is heterogeneous in makeup, and most of it is in paper-based form. This information could provide a baseline of knowledge for biodiversity, but the identifying data, although captured on labels or preserved in field notes or photographs, has neither been digitized nor made widely accessible. There are instances in which the specimens are held in storage for preservation purposes but have not as yet been properly identified and catalogued. In other cases, specimens are stored, but not in a manipulable form because of the diverse collection methods used over time. For example, the American Museum of Natural History (AMNH) has materials such as photographs, field notebooks, maps, paintings, and other physical artifacts from expeditions that took place in the earliest part of the twentieth century. These materials are valuable for study purposes, but until recently they have not been available in digital (interactively analyzable) form.

Tom Moritz, AMNH, described current field collection activities as the "new natural history." Modern collection practices include the use of audio files, remote sensing data, geospatial data, frozen tissue samples, photo-traps, and the capacity to collect DNA sequencing information. The use of such diverse data and data formats creates new concerns with regard to the preservation and integration of information. In addition, the natural history community appears to be moving towards a much more rigorous accountability for the impact of on-site collection practices, so that methods of collection and specific practices must also now be captured and preserved. The architecture of a global biodiversity information system must be constructed to permit integration of these new forms and formats.



Researchers and other professionals have responded to the priority of identifying, naming, and collecting species through attempts to capture knowledge electronically and post it to the Web for widespread access and retrieval.

Indeed, as Dan Janzen, University of Pennsylvania, joked at *MetaDiversity III*, researchers fantasize about the creation of electronic devices from the realm of television science fiction. Janzen said, “I want a tricorder, like Spock in Star Trek. A spaceship lands. An alien steps off and holds a gadget out and points it at various objects around him. The machine identifies it as a plant and tells the alien whether it is edible or dangerous.”

There are amazing, highly efficient new ways in which to capture data, physical material, and other forms of information that researchers seek to make available globally in the greater interests of our planet. The priority is to develop the means to make that data available once it has been captured.

Researchers, scientists, information professionals, and others have quickly moved to create the tools necessary for collecting, aggregating and disseminating the data needed to support knowledge creation and informed decision-making. New data entry tools, new identification tools, new analytical tools for data analysis and data mining, and the standardization of metadata are emerging as the building blocks of a global information system.

### **Capturing Information in the Field**

**Dr. Daniel H. Janzen, Professor of Biology and Thomas G. and Louise E. DiMaura Term Chair, University of Pennsylvania**, has done extensive field work at the Conservacion Guanacaste in northwestern Costa Rica. His approach towards information handling is very different from that of those who work either in museums or information environments like libraries.

Janzen began his discussion of capturing field information by presenting the concept of an “event” – an instance of a field researcher finding an organism in the field. In current research practice, the field parataxonomist identifies and catalogs the occurrence of the event (the finding of the organism) by creating a record, which is subsequently uploaded to the Web so that another individual, perhaps a user in a classroom thousands of miles away, can access it.

In the field of biodiversity, Janzen suggests, an event poses two questions: (1) Why focus on the process of finding and cataloging the organism, and (2) How do we construct an effective process for making that recorded information available. The “why” is readily answered in pointing to the preservation of the richness of our

environment. If wild tropical biodiversity can be used in non-damaging ways by various sectors of resident, national and international society, it has a significantly greater chance of surviving.

The “how” is answered by looking at today’s technology and using it to improve outdated methods. For many years, the richness of the natural world was made available to mass audiences through museum collections (where specimens could be viewed, as well as properly studied, cataloged, and preserved by experts), and through the field guides and reference collections archived in research libraries. These institutions have had the responsibility of curating and protecting human knowledge of natural history as assembled over the past two hundred and fifty years. But, insists Janzen, computers provide an alternative means for capturing and disseminating information for the field scientist today, an alternative that does not require the costly inclusion of museums or libraries. By streamlining the process, by gathering and processing information in the field, scientists and users derive a multitude of benefits:

- Ensured storage of information, regardless of the location of the individual researcher;
- Consistency and interoperability of record-keeping within a specific project as well as in conjunction with other projects;
- Ability to review project records on an on-going basis; and
- Easy transference of information to others with similar interests.

Janzen passionately insists that museum specimen collections and informational databases are *not* where the dissemination of biodiversity knowledge should begin, but rather that dissemination should start at the point of inventory at a survey site. The actual event occurs in the field (seeing a purple caterpillar, stepping on a form of moss, etc.), and the collateral information documenting that event should be captured at the same time. This information includes such field data as the collection date, country, province, locality, sector, and Lambert coordinates, as well as phylogenetic information as to subfamily, order, and species. There may be other forms of data attached to a specific event, such as images, genetic sequences, or other physical elements. All of this collateral information is connected to the event via a unique voucher alphanumeric. The repetition of event documentation in a survey site or as collected over the global distribution of a taxon results in masses of highly particulate information. The challenge is the dissemination and manipulation of that mass of information, the availability of which will accelerate and expand scientific discussion and discovery.

By encouraging the working field scientist to do the information-recording on his or her own, the value-add contributed by museums and libraries to natural history collections may need to change, albeit not disappear.

“When you start producing this stuff on the Web (i.e., remove it from the context of a published article in a journal on a library shelf), obviously the responsibility for preserving the material shifts back to the scientific community, and it is not ready for it. As a working scientist, I’ve always abrogated this to someone else, to publishers and librarians.”

Like so many other researchers with an interest in biodiversity, Janzen actively maintains a highly specialized information resource on the Web, consisting of more than 200,000 dynamic records in a database located at <<http://janzen.sas.upenn.edu>>. As a form of incentive for encouraging field scientists to engage in this form of resource-building, Janzen described how he reduced his operational costs in the field by assembling off-the-shelf software and standard hardware to create this knowledge base.

“In 1995 when we started thinking about doing a massive inventory in Costa Rica, we budgeted approximately \$100 million dollars to do a species mass equal to North America and Canada in an area the size of Philadelphia and surrounding suburbs, Janzen stated. “Thirty percent of that \$100 million was clearly identifiable as computer management and related expense, including personnel, machinery and the software. Today, we can handle almost all of that with off-the-shelf stuff like FileMaker, Excel, HTML editors, etc. It’s not even one percent of that original 1995 budget.”

“The machinery for delivering this highly-structured database is two amateurs, myself and my wife [Dr. Winnie Hallwachs, University of Pennsylvania], buying stuff off the shelf. I’ve never read a computer application instruction manual in my life. But the process for this is really very simple and now kids can even do this.”

The flexibility to correct or modify records is necessary, as a record will incorporate changeable elements in order to accommodate revisions over time. Janzen’s example of such updating was an instance of a locality experiencing a name change. “Before, it was ‘Place name A’ and the records reflected that. But something occurs politically and the name of the town becomes ‘Place name B.’ The existing records must now reflect that change. Otherwise, when people going in looking under ‘A,’ they will never find those instances of occurrence after it becomes place name ‘B.’”

Janzen’s Web site incorporates both static and dynamic pages. “Species pages are a frozen report, equivalent to a page ripped from

a field guide or other synthesis of biodiversity information.” He compared these static pages to plates in a book; he uses .jpeg images as archival masters. These images are produced at a maximum resolution and never alter. All images throughout the database bear unique identifiers that relate to the master voucher.

In discussing modern field research activities within the context of gene sequencing for species identification, Janzen talked about the availability of a device that would enable the researcher to insert an identifying element of an organism, such as a leg from a moth specimen. Then, after a single moment, a single genetic sequence would be generated by this device – at the cost of a single penny. To be able to use sequencing as a form of bar coding would provide a method of documenting events on a survey site that may hold, as in the case of the Guanacaste reserve, more than 100,000 eukaryote sequences. Janzen postulated that he would be able to upload that sequence into some sort of a central database or even some segment of a database held locally or radio-linked to a computer in his field tent. “Within two seconds, it tells you ‘This is a new species’ or ‘This is not a new species.’ If it is already in the database, you ask for all the collateral information that goes with it, either summarized in species pages or in particulate form, record by record.”

“Nature,” Janzen stated, “does not begin in a museum. It begins out there (in the field). We can ask a few questions about many, many things when we do a biodiversity inventory. For each species, get a working name on it, figure out where it is, understand the minimal parts of its natural history, and how do you get it when you want it. Oh, and then Internet it.”

The urgent need to capture as much biodiversity information as possible before it is lost (whether due to human activity, environmental changes, or other factors) has caused the research community to enlist the wider population in activities of cataloging and identifying regional plant and wildlife.

### “Citizen Scientists”

At the *MetaDiversity II* meeting two recommendations emerged:

- Encourage both scientists and members of the civil society – amateurs, ecotourists, birders, gardeners, classrooms, scout troops, fishermen, hunters and life-long learners – to participate in online data collections and to support funding of such efforts. Develop and make available training materials for parataxonomists; and
- Involve the children. Prepare the biodiversity researchers of the future.

In 2003, it is heartening to see how much has been accomplished in response to those recommendations. Citizen-based biodiversity surveys, mentioned frequently at *MetaDiversity III*, use the efforts of novice collectors to identify, count, and evaluate organisms within a specific territory. Dr. Jansen uses the local workforce (parataxonomists) to assist in his inventory in Costa Rica. In the domestic United States, states such as Illinois, North Carolina, and Pennsylvania have established local initiatives, using citizen scientists. For example, on June 27, 2003, a New York City Central Park “BioBlitz” took place for the first time in the park’s 150-year history. Three hundred and fifty citizen and scientist volunteers recorded their findings on lightweight computer notebooks from different manufacturers and using specially designed software created by volunteers from Microsoft. These naturalists identified nearly 850 species within 24 hours, not including microorganisms collected from the lake (see [http://www.usatoday.com/news/science/2003-06-29-creatures\\_x.htm](http://www.usatoday.com/news/science/2003-06-29-creatures_x.htm).)

### **Identification Keys**

Given an increasing reliance upon a novice workforce to participate in essential field activities, it has been necessary to create support tools for these citizen scientists. As an example of such a support tool, it is useful to look at the development of botanic and biological keys.

Bryan Heidorn referred to several projects focusing on the development of digital tools to support the field work of citizen scientists and researchers as examples of how and why the electronic information environment is a better mechanism (than paper) for supporting biodiversity studies.

The ultimate objective in the development of the tools as outlined by Heidorn is to make the biodiversity knowledge available and accessible to workers anywhere and at any time, not just within the physical confines of a library or even on the Web.

### **BIBE**

BIBE (Biological Information Browsing Environment), a project now completed, was a National Science Foundation-supported project, collaboratively done by the Graduate School of Library and Information Science at the University of Illinois, the Illinois Natural History Survey, the Missouri Botanical Garden, the Flora of North America Project at the Harvard Herbarium, and the Illinois Department of Natural Resources. It was initiated in 2000 and completed in 2002.

The objective was to facilitate access to online flora and fauna by both novices and experts through enhanced indexing, searching, and

visualization techniques. According to the project Web site <http://www.biobrowser.org>, “Specific search facility and content will be added to help users with different levels of domain knowledge to identify species, based on the augmentation of professionally developed taxonomic treatments or species descriptions.”

The BIBE project was specifically oriented to the development of software tools that would support the gathering of information by volunteers or “citizen scientists” in biodiversity surveys, during which individuals observe and document organisms within a specific region and within a confined time-frame as noted earlier.

BIBE involved the conversion of scanned page images from the Flora of North America project into XML (eXtensible Markup Language). The software developed and used in BIBE could parse the scanned image representations into appropriate sections of taxonomic descriptive data, understanding what segment of the page was nomenclature information and what sentence or clause offered a leaf description or provided the root structure of a given plant. Additionally, the XML facilitated the automatic indexing and search functionality of the captured pages. Based on the XML, the project team was able to make structured index files of subsections of the document for searching by the user. In addition, they created an online glossary and structured it to provide automated query expansion for enhanced retrieval and other recommended functionalities. All of the tools created were open-source software and were created in accordance with standards from the International Taxonomic Database Working Group. (Further details about the project are available at <http://www.biobrowser.org>.)

Upon completion, BIBE “morphed” into two current projects, OpenKey and TeleNature.

### **OPENKEY**

OpenKey, a project-partnership between the University of Illinois at Urbana-Champaign and the University of North Carolina at Chapel Hill, is focused on the development of additional interactive keys that will help volunteers involved in statewide biodiversity field projects to identify species and gather data that can, in the long-term, be used by others involved in biodiversity studies.

Of primary importance in OpenKey is the development of what are known as polyclave keys. These are keys that allow the user to select from multiple elements, as opposed to dichotomous keys that force users to pick between only two elements before progressing with their search. These polyclave keys will also be error-tolerant in instances where a worker might enter a single element or

characteristic that is incorrect, and they will be supportive of approximate matching and dynamic re-ordering of keys. Critical system attributes of such polyclave keys include data entry interfaces, data storage, and user interfaces for search and retrieval.

Other objectives of the OpenKey project focus on the development of resources (information in both text and image formats) that will be made available to the general public in support of the statewide biodiversity field work.

The tools and resources created by OpenKey will be made available to other libraries and museums. More detailed information about this project is available at <http://www.isrl.uiuc.edu/~openkey/> and at <http://www.ibiblio.org/openkey/> (a site sponsored by the University of North Carolina).

#### **TELENATURE**

TeleNature is an interactive tool designed for use by individuals having varied skill levels in biodiversity studies and who may be working in the field to identify species. It is currently used by students (K-12) in data collection at the local and regional levels, providing both images and text to aid in species identification.

By using wireless devices, such as laptop computers and PDAs, which can support both text-based and live reference services, workers in the field can more efficiently gather the raw data that documents the habitat, location, migration, and other attributes of organisms. Because data from such wireless devices is transmitted to field servers and ultimately to stationary base servers, professional researchers can assist the novice collectors in species identification and documentation when difficulties arise. The interactive aid provided by the professional researcher is also documented and added to the store of knowledge, so that data is continually enriched for use by future workers and volunteers.

TeleNature has been used in conjunction with the Illinois Department of Natural Resources' EcoWatch Network. For more on the work done to develop TeleNature, go to <http://www.isrl.uiuc.edu/~telenature/>.

#### **POLYCLAVE AND DISCOVERLIFE.ORG**

PolyClave and DiscoverLife.org are two additional initiatives in which interactive identification keys are a focus. Developed by the Canadian research community, PolyClave runs on the Delta software originally developed by CSIRO (Australia). Information is available at <http://prod.library.utoronto.ca/polyclave/>.

DiscoverLife.org is a Web site <http://www.discoverlife.org/> that is a biodiversity information resource, aggregating content from a wide variety of botanical gardens, government agencies, non-government organizations and educational institutions. Interactive keys are a part of the many tools available to the user.

#### **Other Data Entry Initiatives**

NatureServe, a not-for-profit organization based in Virginia, is working with scientists to develop a standard for a biodiversity data conservation model. Larry Sugarbacker, Chief Technology Officer for NatureServe, told *MetaDiversity III* participants that NatureServe <http://www.natureserve.org/>, in conjunction with the software company ESRI <http://www.esri.com/>, sponsored a workshop in 2001 in which approximately 20 different organizations came together to discuss a biodiversity data conservation model. The workshop objective was to identify the essential object components of the model so that standards could be developed around it.

*"The advancement of science should not be limited by habit [of print], no matter how productive the habits have been in the past."*

The challenge is to help scientists in the field to collect data in a structured way for future integration into NatureServe datasets. To this purpose, NatureServe has begun the development of new field data collection tools. The technology that they ultimately will use will be either hand-held technology or the new Tablet PCs. This technology will have the ability to collect geospatial data in the field, either directly from images or via a Global Positioning System, and allow scientists to enter tabular data in the field as well. The interesting thing about this initiative is not that they are automating the field data collection process but that they are building a field workstation tool that will permit the generation of field data collection templates from a standard data model.

What is unique about this? When developing a new data collection tool, organizations typically will base it on an existing field data collection protocol *or* develop a protocol that reflects a dataset relevant to the specific survey being performed. The organization then builds a set of paper forms or, if they are technologically savvy, they may completely automate the process. Unfortunately, the NatureServe project does not have the benefit of an existing standardized data model. In the development of their tool, they will take a standardized data model representing the needs of a broad set of organizations and integrate that new model into a field data collections forms development process.

When designing a form, the user will select an object from the database by clicking on it and dragging it down onto the form. This will simplify the process, because a user can add data such as a species name without worrying such things as, “How many characters are involved? How do I represent it? And, importantly, How do I spell it?” The NatureServe application will be able to incorporate a standard data field so that the user can control the data design process at the very point when a research assistant is first thinking about going into the field to collect data.

### **Specimen Collections and Observation Data**

As mentioned previously, collections of specimens in museums, herbaria, research institutions, and other facilities represent a massive body of information, although that information may not yet be in a digital or manipulable form. Biodiversity information and knowledge is widely held by the global community in the forms of specimen collections, observational data, archives and manuscripts, satellite images, sequencing data, film, and bibliographic indices.

Collections extend back in time, in some instances as far back as 300 years. However, without the appropriate identification and cataloging of each item, the value to be derived from the specimen is limited. Specimens may be prepared and stored in a variety of ways, and several preparations from a single specimen are possible (as, for example, when the skin and the skeleton of a specimen are held and recorded as different items within a collection). Specimens can provide information regarding the geographic distribution of a species, and dating may indicate how that distribution has changed over time. Additionally, written field notes, labels, maps, and drawings may be part of the overall body of captured information. As noted earlier, this significant body of information could provide a baseline of knowledge for biodiversity, but the identifying data, although captured on labels or preserved in field notes or photographs, for the most part have neither been digitized nor made widely accessible. In some instances, specimens held by museums may not even be adequately identified and catalogued, due to the lack of sufficient resources. In the past, changes in migration patterns, ranges, and behavior has been difficult to capture. However, the data required to determine such changes may actually be held in the field notes and labels associated with each specimen, but the funds required to analyze the notes are not available.

Observational data, such as that gained from the use of photo-traps, document the presence of a plant or animal at a specific point in time and location. However, in an attempt to preserve on-going biodiversity, no attempt is made to capture any physical artifact. In such instances of data retention, the photograph, audio recording,

remote sensing data, etc., represent the only record for subsequent study by others.

In the past, researchers and scientists cataloged and indexed the current store of knowledge to the best of their ability through traditional printed publications such as monographs. This print format requires scientists to work in a library of physical reference works, amidst hundreds of printed volumes, three or four volumes open in front of them at a time, cross-checking descriptions, characteristics, and other information in attempts to establish accurately new species or document supportive materials regarding changes in patterns, behaviors or migrations. Due to the lack of adequate electronic and digital tools, the research community has no current efficient method to make correlations between segments of knowledge. Nor are the necessary print tools useful when working on biodiversity surveys in the field, because they are too numerous and too cumbersome to be carried along for reference. The indexing and cataloging task is enormous. Libraries and books have failed to resolve adequately the problem inherent in biodiversity information management. A better solution is needed.

“The advancement of science should not be limited by habit [of print], no matter how productive the habits have been in the past,” Bryan Heidorn, University of Illinois, Urbana-Champaign, affirmed. He raised the question, “What other kinds of indexes should we have at our disposal to access other segments of the full descriptive data?”

Heidorn is currently involved in the development of software tools that will support distributed taxonomic description and identification. He stressed that the application of digital technologies is the appropriate next step in the gargantuan task of cataloging and identifying organisms. This task can be best accomplished through the use of interactive tools in conjunction with the ongoing construction of information storage systems that are sufficiently flexible to support the gathering of new knowledge. By more effective aggregation and dissemination of information via these systems, field workers can accelerate the accurate identification of specimens, thus equally accelerating the accumulation of a global biodiversity knowledge base. He listed a number of enhanced functionalities that would support researchers within an electronic knowledge environment:

- Improved indexing
- Query expansion (via synonymies, thesauri, etc.)
- Vocabulary switching (in support of cross-disciplinary and interdisciplinary work)
- Online glossaries providing hyperlinked, in-context definitions

- Interactive keys (In the context of biodiversity, keys are individual characteristics of a specimen for classification purposes. In software, these keys become filters applicable to underlying datasets, allowing workers to select multiple descriptive elements to facilitate appropriate identification and classification of a particular specimen. Interactive keys specifically enable workers to input specimen characteristics in any order rather than the enforced order of their printed key counterparts)
- Publishing tools
- Synchronization between descriptive data and characteristic key data
- Dynamic key generation
- Progressive publishing
- Extended space for descriptions and images
- Print on demand capability according to region, family, habitat, habit, etc.

What is essential, therefore, according to **Dr. Tom Moritz, Boeschenstein Director, Library Services, American Museum of Natural History (AMNH), (NY)** is a more rigorous, analytical way of thinking about the data, information, and knowledge for which the research, library, and museum communities are responsible. Equally essential is the need to transfer non-digital information (legacy data) cost-effectively and efficiently to the digital environment. A standards-based approach towards capturing the key data elements common across specimens and collections in a variety of physical and virtual formats will allow practitioners, researchers, and curators to transfer effectively information related to these specimens and collections into an environment where it will be fully searchable and accessible by all.

Use of the Darwin Core standard

<<http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV2>> – a modified standard for the natural history community based on the elements of the Dublin Core <<http://dublincore.org>> metadata standard – is one way in which an efficient information transfer might be handled. According to Moritz, use of standards is essential when migrating hundreds of years of legacy data into the digital environment. “If we put the effort into making a rigorous analysis,” said Moritz, “then we stand a far better chance of making decisions that are well-informed and extendible into the future.”

It was noted, however, that standards developers must not create barriers to adoption by the larger community simply by “overdeveloping” the standards.

Because of the scale of the task of migrating large amounts of data to a digital format, those involved in the process must be parsimonious, efficient, and optimal in their selection of the common elements that will be applied to the broadest set of data formats: what elements should be captured, where those elements are explicitly stated, and when those elements may be inferred from internal knowledge regarding the collection. Cost considerations are critical aspects as well.

Moritz cautioned, “When we make decisions about mark-up, there’s a cost-vector implied when we try to scale this problem out over the whole domain, the whole body of knowledge, the universe of information.”

### **Case Study Illustrative of Challenges**

Moritz discussed the practicalities of transforming legacy information into digital form, using as an example the AMNH project entitled the “Congo Expedition” <<http://diglib1.amnh.org>> having collected materials dating from May 1909 to November 1915. The project involved the digitization of 8 volumes of field notebooks, the creation of digital photographs of 4,000 anthropological artifacts, and the digitization of 2,200 photographs taken by the principal investigator, Herbert Lang, and 98 water color sketches completed by his assistant, James Chapin. In addition, the AMNH created digital editions of 160 publications based on the findings of the Congo expedition. All of this content was made searchable for the Web.

Moritz also reviewed the efforts involved in capturing the identifying text from labels, envelopes, and similar items in order to make the information accessible online. This was accomplished by creating metadata using cost-effective, known best practices.

Referring to the MARC standard <<http://loc.gov/marc>>, used by the library community for the representation of bibliographic and related information in machine-readable form, Moritz stated that the cost associated with the creation of a MARC record was \$13.00, a cost that will not scale as an economic option. “The fact that we have a body of these records already in hand for many scientific publications is an advantage because it permits us to translate and to inherit metadata from other standards. The AMNH developed a mediated Dublin Core standard for use in creating records for the Congo Project. The standard was less expensive, but not in and of itself a more efficient solution.”

Handling native or vernacular metadata is also challenging. As an example, Moritz displayed an artifact, an envelope containing a

photographic negative with a handwritten caption, “Leopard, male, shot by a pygmy, with an arrow through the heart. The two men are Pygmies.” How does a curator with a minimal amount of mediation turn that descriptive caption into a valuable and useful piece of information? Readily determined information such as the scientific name of the leopard or the name of the photographer may not be as problematic, as it may be inferred from knowledge of the specific collection. Moritz advocates the use of appropriate standards, such as the Resource Descriptive Framework <<http://www.w3.org/RDF>> and the Dublin Core. By working within those parameters, it becomes possible for staff to transfer identifying text and other information to an electronic record, minimizing the cost of transfer while preserving and making searchable the relevant information.

Given the uneven development across the world in natural history institutions, it is appropriate to consider the questions of the most logical way in which to develop the desired information architecture and, practically speaking, who can afford to do it and at what level. A great deal of data exists that has not been captured, converted, and made available.

At one point, Gary Rosenberg of the National Academy of Science in Philadelphia asked Moritz to comment on the role of preserving the verbatim data. Could tools be developed that might help the community to retain the original information as presented, incorporating natural language into the parsed data?

Moritz responded, “I believe one of the fundamental rules of processing is that you maintain the original record. At the AMNH, we’re preserving both the transcriptions and the actual images when we are dealing with digitization of field notes. “

With regard to the tools that might help, Moritz seemed hopeful. “In connection with analyzing the original language of geographic descriptions, particularly in legacy data, there are folks working and finding probabilistic ways of expressing a polygon based on that information, and those rules can be very clearly specified. For example, if you have range and bearing from a known place (such as ‘We’re three and one-half miles north by northeast of Place Name X in the Belgian Congo’), you can come up with some rules that specify to some degree of probability what polygon the original researcher was probably in. That work is being advanced. It’s one way that we can take that native raw data, specify rules, and give it an expression with some level of confidence in it.”

## Discussing The Value of Metadata

Within the context of Moritz’s presentation of creating metadata for legacy collections, one audience participant posed the simple question, “When does data become metadata?”

Moritz’s response was that metadata is really just an operational issue. “If I can expose my information and data in some way, and it can be operated on in a Web environment using some of the things we’re discussing here, it does not really have to be centralized and it doesn’t have to be presented formally as metadata. If I have it in XML, if it is marked up according to a common standard, and if that rigorous analysis that I’ve referred to is accurate, then it is really a question of whether the custodian of the data is capable of (a) managing their own data and (b) consistently exposing it. Or is it more efficient for them to give it to a central repository? It’s more of a practical, operational choice.”

Heidorn enlarged upon the response, offering: “Whether a piece of data is metadata or just data is dependent on how it is used, rather than on what it is. If you are using the data as a level of indirection to get to other data, then it is metadata. But frequently that data itself is useful without actually following it out to the other data. The user, at the point of need, determines whether it is data or metadata.”

In a related but separate discussion, **Bruce Westcott, an independent consultant** attending *MetaDiversity III*, highlighted the value of metadata in two points:

1. The creator or custodian of the data benefits because metadata maintain the value of the dataset, facilitating its continued use over time.
2. Users benefit because metadata allow them to find and use geospatial data. In particular, if metadata are created according to the federal metadata standard and are contributed to a National Spatial Data Infrastructure Clearinghouse, it becomes possible for other users to access and retrieve this information and its related data effectively.

Metadata that meet a standard are easier to share, to search, and to use.

Metadata largely address issues of quality, reliability and authenticity. They protect the long-term value of the asset represented by these datasets by enhancing them for use in decision-making. Metadata ensure that users can determine if the quality, accuracy, and currency of the data is sufficient for the purpose to which it is being applied.

“In the world of spatial metadata,” said Westcott, “it is problematic to define metadata simply as ‘data about data.’ In fact, to provide a

meaningful description of geospatial data, it is far more useful to specify the actual elements that are captured. Take as example the definition on Geomatics from the International Organization for Standardization (ISO). This definition lists metadata as the content elements dealing with “the identification, the extent, the quality, the spatial and temporal schema, spatial reference and distribution of digital geographic data. The value of standardized metadata is that it is easiest to share, to search, and to use.”

### **Geospatial Data Collections**

Geospatial data are of central importance in understanding biodiversity. The description of a dataset (or metadata) has become increasingly important for locating and accessing information of all kinds. A standardized conceptual schema for geographic information metadata will increase the ability of geographic information created for one application to be accessed and evaluated for use in another application.

Just as the AMNH is focused on the capture and use of metadata for both legacy and new data, the Florida Marine Research Institute (FMRI) is similarly concerned with increasing the functionality of the data that they have been collecting for the past 55 years. The Institute <<http://floridamarine.org>> sponsors more than 130 research projects at its 12 field stations, projects involving extensive collections of biological, chemical, physical, and geographical information. In particular, the Institute has worked extensively in the collection, curation, and dissemination of the coastal and marine geospatial data associated with the 8,400 miles of coastal areas in the state of Florida.

One of the significant contributions made by the FMRI is the Coastal and Marine Resource Assessment program (CAMRA). This program provides analytical tools, including sophisticated mapping technology, that support state policy decision-makers as well as the scientific community. The data collected by CAMRA date back to the early 1980s, and a key objective for FMRI has been to ensure the long-term accessibility of the data for use in resource assessment and preservation.

To facilitate the research community’s use of the extensive data assembled over the course of several decades, **Jill Trubey, Metadata Coordinator, Florida Fish & Wildlife Commission, Florida Marine Research Institute**, has a key role in ensuring that appropriate metadata are applied methodically and uniformly. According to Trubey, “One of the FMRI’s long term goals is to provide high-quality, uniform data for all datasets generated within FMRI.” Managing and providing high-quality data is critical to the success of the FMRI mission to protect, conserve, and manage Florida’s marine and coastal resources.

Metadata that conform to the Federal Geographic Data Committee (FGDC, at <<http://fgdc.gov>> standard are the basis of the National Geospatial Data Clearinghouse, a distributed online catalog of digital spatial data <<http://clearinghouse1.fgdc.gov>>. The use of FGDC-compliant metadata facilitates the understanding of diverse datasets throughout the research community by describing them in a way that emphasizes aspects that are common among them. In the mid 1990’s, the NBII and the FGDC developed standards for metadata use in the United States. However, international standards have not yet been finalized. As Trubey and others reflected over the course of *MetaDiversity III*, while standards may be critical to successful integration of information systems, they are frequently too complex and laborious to implement.

Given the importance of large datasets across many scientific disciplines, it may be surprising that organizational attitudes towards the application of metadata can be somewhat cavalier. Without a firm grasp as to how metadata ensure the continued utility and value of an individual’s lifetime research, organizational staff may be reluctant to devote the time and resources required for metadata creation. Trubey was able to convince FMRI management of the value of metadata, but admits that it was one of the most difficult aspects of the project. The role that she plays as Metadata Coordinator is critical to the on-going communication process with managers, a process that strengthens organizational commitment to the metadata project. The methodical and uniform application of metadata requires a strong individual commitment, including expertise on a day-to-day basis with regard to international standards, metadata software applications, and information infrastructural requirements.

Finding metadata software that was compatible with their internal implementations of off-the-shelf database software from Microsoft and/or Oracle was a core consideration. According to Trubey, FMRI selected the Spatial Metadata Management System (SMMS) software <<http://imgs.intergraph.com/smms>> as being compatible with their needs as well as compliant with both FGDC and NBII metadata standards. The well-designed software enables staff to create keyword lists, to use naming conventions and standard responses to specific fields, and allows for the use of templates. Trubey spoke well of the SMMS software in facilitating their efforts, given that the end result is consistency across FMRI datasets. Documentation was provided to the staff to walk individuals through the necessary processes. A hierarchy of internal department liaisons and technicians also ensured that staff understood how the metadata was to be created. Hands-on training, day-to-day contact, and collaborative email trouble-shooting bolstered efforts to educate



FMRI's more than 400 researchers. The Metadata Coordinator bore the final responsibility for quality control before the final upload of material to the organization's intranet for use by FMRI scientific community.

According to Trubey, "Documentation of many years of data will not happen overnight, but a consistent, methodical approach will go a long way towards getting the job done. FMRI's plan is in place to accomplish this. The legacy data that we as a scientific community have spent millions of dollars to acquire is being documented. We started with core data first (more than 450 of FMRI's data layers were documented as of April 2002) and are now working to document all new and current projects. We will then strive to capture metadata for our historical data. This task will be particularly arduous because much of the information is on loose data sheets or poorly organized in notebooks, and many project principal investigators have moved on to other jobs or retired. The effort of capturing these metadata will be time consuming and costly, but the investment will be worthwhile."

Trubey's contribution underlined the importance of several practical considerations in building interoperable research tools:

- Familiarity with national standards such as those used by NBII and FGDC will make it easier for both regional and national agencies to provide access to enriched datasets for analytical consideration and for use in environmental policy decision-making. As international standards emerge and the global community moves forward in constructing open-access information networks, organizations must gain expertise in implementing those as well.
- Consistency of application and uniformity in constructing metadata is central to success. Organizations and agencies will be well-served by appointing a full-time employee as an authoritative overseer for the tasks of quality control assurance in order to achieve such uniformity.
- A well-coordinated effort at application of metadata to key resources will ensure the longevity of collected research datasets and enhance its utility in the future.

FMRI treats their data as a serious asset. They have a lifecycle maintenance approach to sustaining and increasing the value of that asset.

### **Analytical Tools from Legacy Data**

Institutions, working at the local and regional levels to document the biodiversity within a specific geographical area, frequently assemble

information resources drawn from both literature and specimen collections. Such resources are created to support the efforts of scientists, researchers, and policy-makers to understand better the impact of human society on the environment.

**Steven Clemants, Vice President for Science, Brooklyn Botanic Garden (NY)**, provided an excellent analysis of how such database tools facilitate analysis for better decision-making. The New York Metropolitan Flora (NYMF) project <<http://www.bbg.org/sci/nymf>> has begun to document the approximately 3,000 species to be found in the 30 counties making up the New York Metropolitan region, encompassing approximately 7,650 square miles. Approximately 7% of the total U.S. population resides in that geographic space, predominantly in urban areas.

The Brooklyn Botanic Garden, founded in 1910, is committed to engaging in plant science research in order to expand human knowledge of plants and to disseminate the results of that research to other science professionals and members of the general public. While the plants of the northeastern United States have been largely documented over the course of the past 250 years beginning in 1743, the changes in the region generated by the importation of non-native or invasive species have not been as well documented. The NYMF project, with the intent of being a comprehensive study, documents the presence of both native and non-native vascular plants within the region and presents the information to the general public via the Web.

The NYMF project, also known as the Ailanthus database, has a number of information components, including descriptions, keys, nomenclature, phenology (seasonal responses of a plant or animal to climactic changes), distribution, and materials drawn from the scientific literature. The structure of its development has been planned so that each database segment, reflecting a significant segment of documented research (such as "All Woody Plants"), will be completed before the next segment is begun. Clemants noted that the Garden has completed the Woody Plants and will next move to the documentation of aquatic and wetland plants.

Each entry for the 450 species of woody plants has images, distribution maps, and technical and non-technical information. Non-technical information includes common names, field identification descriptions, and the various uses of the plant. Technical information includes nomenclature and in-depth descriptive materials – including phenology, habitat, distribution, rarity status, species biology – and literature references. Literature references are drawn from a bibliographic database, searchable by genus, by locality in the region, and by a variety of other attributes.

The distribution maps provided as a part of the NYMF do not use latitude and longitudinal coordinates. Rather, Clemants has chosen to map locations by using Universe Transverse Mercator grids (UTM). These blocks consist of a 5 kilometer by 5 kilometer square area. The value of using this grid is that it makes it much quicker to geocode the data and makes it usable. If deemed necessary, latitude and longitude data may be added at a later date.

The Ailanthus database contains 225,000 records of which 60,000 are vouchered records from 12 important herbaria in the Northeast region. The remaining 165,000 records are literature or observational records. All are searchable via 9,000 scientific names. The UTM blocks cover 18,500 localities.

To facilitate the construction of the database, Clemants and his team culled specific information from the collections of other herbaria: the name of the plant, the area in which it was sighted, and the year in which it was sighted. Another page was generated with the rest of the relevant information with an electronic pointer to the primary collector of the data (whether by specimen-collection number or accession number in the herbarium that housed the original record) or with a pointer to the bibliographic information from the database. "In this way," said Clemants, "with a very small group of people, we've been able to amass a very great deal of data."

The strength of the NYMF lies in the manner in which the data has been coded to permit manipulation for further analysis. Clemants used a technique borrowed from the Flora of Great Britain project known as a change index. That index provides an indication of how the range of a species has contracted or expanded between an early era (such as 1900-1950) and a later era (1950-2000). According to Clemants, "Some of the things we're now able to realize is how whole groups of native plants – for instance, blueberries – are declining in the New York area over the last 100 years, and we can start to ask why that is. Any number of things may have an impact, and this enables us to see that pattern and start to ask why this is happening in our environment."

The analytical work that the NYMF project supports has allowed the Brooklyn Botanic Garden (NY) to establish close working relationships with the Long Island Weed Management Area and with the Delaware River Invasive Plant Partnership in order to control invasive species. The data provided by the NYMF project allows policy-makers to identify problems, ask which problems are the most serious, and set appropriate priorities to protect the environment.

### **Biodiversity Analysis Decision Support Tool**

Another example of a tool for the analysis of biodiversity information is from the organization, NatureServe.

Those who knew NatureServe as part of The Nature Conservancy, or as the Association for Biodiversity Information, might have thought of the organization as one that collects, manages, and distributes data. But it is important to know that NatureServe has a broader mission. Currently, they are developing decision support tools so that they can not only collect and distribute data, but also assist users to incorporate those data into critical decision processes.

*"While we are all scientists and we want to be able to represent the data in its accurate and most pure form, the reality is that we need to get our data represented in a way that is useful to decision making processes and can be compared with other data resources."*

According to **Larry Sugarbaker, Vice President and Chief Information Officer, NatureServe**, "While we are all scientists and we want to be able to represent the data in its accurate and most pure form, the reality is that we need to get our data represented in a way that is useful to decision-making processes and can be compared with other data resources."

Sugarbaker used as an example of their efforts in developing analytical tools a recently completed pilot study done by NatureServe in service to Napa County, California.

The first thing NatureServe needed to accomplish for this project was to assemble a customized biodiversity database. It was in this initial step that they faced significant challenges. NatureServe has accumulated approximately 750,000 observations of range and endangered plants and animals across the US. They have access to extensive vegetation classification data and ecological systems data. But, as they quickly came to realize, it is often difficult to know who owns all of the data required to make sound conservation decisions.

The first challenge was to identify where such datasets existed. The second was to re-format the data so that their system could accept and use the data. They found that the amount of data that is readily accessible with high-quality metadata is relatively small. As a result, a vast quantity of potentially relevant data needs to be identified, gathered, and made accessible in order to make informed scientific decisions.

For the Napa Country project, NatureServe obtained data from a handful of organizations, including the California Native Plant Society, the California Fish and Game Department, and the Department of

Forestry. Most of the datasets were structured and provided information on such points of interest as northern spotted owl nesting sites, locations of hardwoods and redwoods in Napa County, etc., as well as NatureServe's elements occurrence data. The location information from NatureServe, and similar data gathered from a number of the other organizations, exists as data points on the ground, locations where a survey has been done and where a particular species has been determined to exist. NatureServe's objective was to transform that data into a characterization and description of the landscape for that particular species – the location of the habitat and the points where observation of the species would most likely occur. NatureServe used existing models and developed new models in order to integrate information about known locations with vegetation communities. The integrated information was then applied to a new model to predict the potential range of a particular species.

The Napa project covered approximately 120 species of interest in that county, and Sugarbaker showed a visualization of what one element value layer (data layer or one species layer) might look like in Napa County. Areas displayed in red on a map were the result of predicting the range of the yellow-legged frog from observational data of where that species actually occurred. The visualization was created using a predictive model based upon highly variable vegetation data. It was essential to know the lineage of every dataset that was used with the predictive model. Users would be making critical decisions based upon these data visualizations that could be synthesized from as many as a hundred different datasets. Analysts had to provide the user with information specifying the authenticity and reliability of the data so that they could estimate the potential accuracy – and ultimate value – of the predictions.

To enhance the value of the analytical tool, NatureServe needed two data layers, one that described the quality of the habitat and one that described their confidence in that prediction.

“In the case of this California project,” said Sugarbaker, “we merged 120 different datasets into a single biodiversity value layer. We needed to be able to characterize that value layer and represent the data relative to its value for conservation. In this particular case, we wanted to identify the location of the highest value conservation lands, and then generate a map showing high value land associated with the aggregation of the biodiversity.” A policy-maker may want to identify areas that are most relevant for protection or gauge how a conservation policy can provide maximum protection with the funds available. This requires knowing several other things, including the location of existing protected areas (not all state or federal land is protected); therefore NatureServe developed classification schema for characterizing that information as well.

When considering possible land preservation, the creation of conservation easements, or even the outright purchase of land for conservation purposes, a major factor to be considered is the location of the least expensive land relative to the highest biodiversity area(s). Tools, such as those created by NatureServe, need to facilitate decision-making by allowing the user to visualize the landscape relative to that factor. In the NatureServe example, a map will display both the most valuable lands (highest biodiversity habitat on a per acre basis), and the lowest loss lands (lowest biodiversity habitat per acre).

In running a scenario for the identification of such land parcels in Napa County, they found that if they wanted to protect viable populations of all identified 120 species, it would take approximately \$90 billion dollars. However, for \$1.7 billion dollars, viable occurrences of a fairly high percentage of the species could be protected. These are concrete numbers that had been previously unavailable to the Napa Valley conservationists and, according to Sugarbaker, they were ecstatic to be able to integrate such information into their decision-making process.

Application providers must be able to create tools that assist in visualizing biodiversity on the landscape, making it real for all those concerned with the protection of these critical areas. There is interest in offering a series of reporting protocols in order to provide feedback to planners so that decision-makers can examine and evaluate a variety of possible protection strategies, and better understand how close (or far away) they are from achieving their conservation objectives.

### **Merging Existing Databases**

**Gary Rosenberg, Academy of Natural Sciences, Associate Curator, Academy of Natural Sciences, Philadelphia, PA**, addressed some of the difficulties that arise when creating a cohesive entity from several disparate resources. He spoke on an international effort to meld several individual databases into the Ocean Biogeographic Information System (OBIS) Indo-Pacific Molluscan Database. The end result of this international effort may be accessed at <<http://www.iobis.org/OBISPortal>> and at: <<http://data.acnatsci.org/obis>>.

The service provides the user access to a full taxonomic schema of names for relevant organisms, maps that provide distribution and migration ranges, fielded searchable data (including such appropriate aspects as depth, habitat, and feeding mode), and access to citations for the published literature.

The institutions joining together to create the system included the National Academy of Science (Philadelphia), the Australian Museum (Sydney), the Museum Nationale d'Histoire Naturelle (Paris), and the California Academy of Sciences (San Francisco).

Among the challenges faced were the redesign of data entry interfaces in order to speed the input of data, the maintenance required to handle data from three different electronic resources, and the use of one of the three databases as a mechanism for updating and error checking.

He stated that his experience has brought him to the following conclusions:

- Developing authoritative lists of names requires large-scale funding.
- Lists of names are not only tools for research, but are also primary data.
- Diversity of data makes convergence difficult.
- Interoperability needs to go far beyond names and geography.
- Synthesis of knowledge may be the rate-limiting aspect in the discovery of diversity.

On the basis of those conclusions, it is appropriate to review the role of standards in bringing a variety of tools and databases together in a single knowledge environment.

### **The Role of Standards: Historical and International**

Spatial metadata was first standardized in the United States at the government level through executive order in 1994. The standard was revised in 1998 (see <<http://www.fgdc.gov/metadata/contstan.html>> and then extended for use with biological resources in 1999 (see <[http://www.fgdc.gov/standards/status/sub5\\_2.html](http://www.fgdc.gov/standards/status/sub5_2.html)>). The key relationship to note is that the FGDC works with the National Committee for Information Technology Standards – Subgroup L1 (NCITS-L1), the latter representing the United States on ISO Technical Committee 211 (the Geomatics Committee). The standard currently known as ISO-19115 DIS was published in August 2001, and a series of implementation steps for this standard will be taken in North America during 2003.

**Bruce Westcott, *Spatial Metadata Consulting***, stated that FGDC will endorse the ISO standard (see <<http://www.fgdc.gov/metadata/whatsnew/fgdciso.html>>, and he urges those interested in biodiversity information standards to monitor the American National Standards Institute. That organization will be responsible for the coordination of metadata and other spatial data information

standards. The development of such standards will no longer be the sole responsibility of the government; instead, there will be national consensus standards.

One key feature of ISO-19115 is a Unified Modeling Language (UML), a highly structured way of describing the metadata. Data dictionary tables define the content of spatial metadata, but the UML is the way that it is formally defined. It describes the relationships expressed in Standard Modeling Language. It has specific provisions for multi-lingual capabilities, largely based upon extensive use of numeric code lists that enable the multi-lingual translations. There are explicit provisions of multiple profiles so that as the standard is applied to species or biological data or remote sensing data, the metadata standard can be extended to the many elements specific to each data type. Core elements are defined, and this definition is well integrated with other ISO information standards.

The ISO standard is more robust than that developed by the FGDC. The terminology differs from that of both the FGDC and the NBII metadata standards in such areas as element names, definitions, and obligations. However, the proposed ISO standard does address some recognized deficiencies in the FGDC standard.

The biggest problem with the FGDC metadata standard has been that it does not conform to any particular data model, and most of the content elements are free text. The ISO standard addresses this specific issue to a great extent. It supports multi-lingual and multi-cultural aspects. From an application perspective (whether that of informatics, library science, or geographic information systems) the key point is that ISO metadata is a structured entity. It provides the real capability to integrate metadata with individual databases and to build applications that draw upon the metadata as well as the data itself. For example, in the area of remote sensing data, there is generally a field for the percentage of cloud cover. This is a well-defined numeric field that can be implemented in different ways. Its relationship to other metadata elements has been well defined so that applications can easily be built that use it as a filter, whereas with the FGDC standard, application providers were largely limited to string search capabilities.

A number of business rules are built into ISO metadata. Core elements can, in many cases, be extracted from the geospatial databases themselves, and institutional or technical parameters or default values can be provided for applications so that the metadata can be auto-populated. Westcott encouraged those addressing metadata planning to become familiar with the ISO core elements and how they are structured. Adopters should consider whether or not they are addressing those core requirements in their current metadata holdings.

The ISO standard also provides for implementation of metadata at multiple levels of granularity. Westcott again referred to remote sensing imagery: “You may have many metadata content elements that are common to a whole series of images, so you can declare metadata at a dataset series sub-level. But we also have many clients who are interested in metadata that are just as specific to particular features or attributes or sets of features in a given dataset so that you can move your metadata down into a relational structure to describe your data.”

With regard to profiles and extensions, data providers have explicit core metadata components in the ISO standard. Providers have a comprehensive profile, but they also have the opportunity and perhaps the need to build a community profile. Said Westcott, “We actually have to have a U.S. profile before we can even implement ISO metadata, so that we standardize things across our domestic usage. Something that’s as fundamental as a basic language and character set for the metadata we cover in profile.”

Many other ISO standards are included by reference. Partly because of the intertwining of other standards, the ISO standard has been adopted, but a project is underway that will develop an XML schema. This will be an implementation model that developers can utilize in order to build on metadata, and one that can be used to validate metadata records – a mechanism that is non-existent today.

### **OpenURL Standard**

Just as the biodiversity community developed the Darwin Core as a modified metadata standard based on the Dublin Core Standard in order to meet their specific information needs, it may be appropriate for that same community to investigate the potential use of a more recently developed standard, the Open URL

<<http://www.niso.org/standards/resources/OpenURL-release.html>>.

The OpenURL emerged from a community of content providers, information professionals and vendors of library information systems. Originally the brain child of Herbert Van de Sompel, Los Alamos National Labs

<<http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html>

>, OpenURL is a method for packaging electronic metadata in such a way that a computer network is able to match users to a variety of digital resources, regardless of where those resources are located, according to authenticated rights of access. Currently a NISO (National Information Standards Organization, <<http://www.niso.org>> standard open for trial to the public, the methodology is in use in a significant number of research libraries in the United States in order to facilitate user access to licensed proprietary content and other materials across distributed networks on the Web.

**Eric Hellman, President, Openly Informatics, Inc.**, a member of the NISO OpenURL Standardization Committee and a physicist by training, remarked that some of the most valuable insights into a problem within one discipline can come from a completely different field of research when anomalies are observed. For example, a particular sequence of DNA in an animal genome may be identical to a particular sequence of DNA in a plant genome. While botanists may understand what the purpose of that sequence may be when it occurs in a plant, it could conceivably open up new avenues of investigation for zoologists trying to pinpoint the role of that sequence in the animal. Linking between those types of information, whether in the published journal literature or in a genome database, is central to identifying those anomalies and pursuing understanding of the relationships in the natural world.

Digital libraries house large amounts of content from a broad range of information providers. These electronic holdings are diverse and can include such items as online public access catalogs, proprietary aggregations of content from single or multiple publishers, and Web-accessible resources or archives. Libraries may own or have licensed the materials or rights of access at great expense, so they are intent on ensuring that local communities of users can easily access and use that content to achieve their research objectives. Prior to the development of OpenURL, one of the frustrations experienced by both users and librarians was the laborious path users had to take to access content. Access required multiple authentication processes due to the failure of networked systems to direct users appropriately to content for which the institution had authorized rights of access. By developing hardware and protocols that allow systems to recognize and direct users to appropriate content via the OpenURL mechanism, this frustration has been greatly reduced in digital information environments on campuses worldwide.

In the context of published literature, the process works fairly smoothly. A user reads an article in an online database and notes a specific reference or citation to another article that might be useful. The obvious functionality desirable here would be a hypertext link between the two documents. The problem faced when implementing this functionality lies in the computer system’s ability (or inability) to recognize (a) what item the user is looking for, (b) where the referenced item resides electronically, and (c) the access rights of that user to the desired item. If the system understands those aspects, the user will be able to access the information sought with a minimal number of clicks.

In the library, a link server stands between one source of content (the original article being read by the user) and the targeted item

(the abstract or full text referenced by the citation in the article). A link server is a software agent that understands the metadata captured within the OpenURL and, in addition, knows the resources available to the user. In this context, the system recognizes the library's valid subscriptions to online services and electronic journals and, when queried by the user clicking on the hypertext link, it responds back to the user with the various channels of access to the requisite information held by the library.

An OpenURL mechanism contains two halves. The first is the server component or the *resolver* (the address of the link server, i.e., the domain name of the library's server such as library1.amnh.org). The second half of the OpenURL is called the query (or *referent*) which is the descriptive information or metadata contained in the OpenURL. This specifies the parameters of the content that the user is trying to access (e.g., the journal title, volume and issue, author name, and similar data that uniquely identify an item). As a URL may contain up to 2,000 bytes of information, there is adequate space for a relatively detailed set of descriptors to be captured for use by the server component in processing the query. Many major content providers in the publishing world support the use of the OpenURL mechanism, and deployment within the North American library community is extensive.

The application of the OpenURL for biodiversity information resources could be fully implemented if the community fully adopts the use of a metadata standard such as the Darwin Core.

The version of the standard currently undergoing trial is Version 1.1. In order to accommodate the widest possible use of OpenURL in collections that may exist outside of the traditional library, the standard includes recommendations for a registry, a mechanism that would allow communities outside the library to channel metadata into the OpenURL framework to create links to content and resources outside the library's purview. Hellman offered an example of an article about an organism such as the banana slug that might contain a link to a record in the *Zoological Record*, a longstanding authoritative reference for that discipline. The OpenURL link would allow the user to navigate to the *Zoological Record* offered by the database provider, Cambridge Scientific Abstracts (CSA), to a genomic database, or to a specimen collection database similar to the one created by Daniel Janzen in the field. In the example used by Hellman, the authoritative body responsible for the aggregation of content was the library, but it might just as easily have been an authority such as the NBII or GBIF.

Tom Moritz, AMNH, affirmed the potential for the use of the OpenURL standard in the biodiversity community. "We, the

American Museum of Natural History, have all of our scientific publications from January 2000 forward available in the BioOne system [a full text content provider; see <<http://www.bioone.org>>]. We already have links for every scientific name in those publications. Right now, the only link is going to the Integrated Taxonomic Information System <<http://www.itis.usda.gov>> system, just to go back and try to expand that naming system, but the example you use is exactly what we're looking for, in terms of expanding that link from a scientific name. It's a model that would serve the entire community."

Responded Hellman, "BioOne is an example of a resource provider that has already physically adopted the OpenURL syntax in their linking system, so it would be an easy thing to move forward."

Time was set aside at the *MetaDiversity III* conference to discuss issues and concerns pertaining to development and use of standards. The notes and recommendations from that discussion group appear in the final segment of this report.

### Repositories and Portals

Returning to the concepts behind GBIF, once standardized metadata has been appropriately added to content, the next concern is how best to disseminate that content over global networks and interconnected systems.

Tom Moritz advocated an approach that involves further development of the Semantic Web, by approaching information from an ontological standpoint. Many researchers and institutions have moved materials into the digital environment simply because there was an immediate need, but it may now be more appropriate to stand back and reconsider, again in the rigorous fashion recommended by Moritz, what the best means for accomplishing an integrated information environment might require. Moritz referred to work done by Tim Berners-Lee in defining "ontology" in the digital environment: "Collections of statements written in a language like RDF that defines the relationship between concepts and specifies logical rules for reasoning about them."<sup>1</sup>

Some institutions will have the expertise and the resources required to create these natural history digital repositories and other resources. But there is a very unequal allocation of resources, technical skills and support throughout the community of museums. The questions are, who can afford to do the full-blown very sophisticated development of their data, and who will have to rely on the other members of the community to do this for them?

<sup>1</sup> T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, May 17, 2001. <<http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.>>

The Semantic Web is still some years away in terms of development, but there are practical online access points currently available that facilitate access by the public. One such access point to aggregated content has been created by the NBII.

The award-winning *My.NBII.Gov* portal created by NBII <<http://my.nbii.gov>> is of value to the wider biodiversity community as a channel for bringing a range of services and content to the desktops of resource managers, scientists, educators and students. **Michael T. Frame, Technology Research and Development Director for NBII, USGS**, provided an overview of the service.

In the late 1990s, it became clear to those working with the NBII that there needed to be a better way to organize, access, and present biodiversity information to a widely diverse audience. Ranging as it did across the general public and K-12 educational communities through the upper levels of the scientific and research communities, as well as encompassing policy- and decision-making bodies, it was clear from the beginning that any service provided would need to be able to deliver heterogeneous content in a variety of formats while also providing ways for users to interact with that content through such capabilities as customization, personalization, and Web access.

The appropriate response appeared to be the construction of an information portal. The NBII is an electronic gateway to biological data and information maintained by federal, state and local government agencies, private sector organizations, and other partners around the nation and the world. The construction of the *My.NBII.Gov* portal site was launched in 2001, in conjunction with similar work being done by the U.S. Department of Defense, following a period of evaluation of various vendors and experimental prototyping. The portal product finally selected for NBII was from Plumtree <<http://www.plumtree.com>>. The initial release to the public was in February 2002.

The benefits anticipated from the *My.NBII.Gov* portal were threefold: (1) the system would enable research via scientific collaboration, integration of data, support for peer-review, and data analysis; (2) it would support communication and collaborative efforts between remote offices that were part of the NBII Node Network; and (3) it would facilitate delivery of information to the general public and to the educational community. These planned benefits have become a reality as the portal supports the work of hundreds of employees in multiple facilities nationwide.

At present, the *My.NBII.Gov* site has in excess of 100,000 documents available to users, between 200-300 “gadgets” or Web services, 32 communities of practice, and 24 publications with content accessible via the Web.

At the same time, aware of the need to protect sensitive resources dedicated to specific audiences, the system deployed by the NBII includes impressive portal security. “The portal’s underlying infrastructure allows us some extensive security all the way down to the document level,” said Mike Frame. Users and groups may see specific categories of documents or restricted links according to the authentication profiles. The crawling agents use different profiles to scan data sources, and links are configured to reflect the security of that primary data source. User access to the various gadgets and Web service applications are subject to appropriate authentication.

Of particular value to users are the “gadgets” made available through the NBII service. Every user has access to gadgets that facilitate the integration of content or applications into their customized page. These gadgets integrate content ranging in scope from ordinary, familiar application such as pinpointing weather by zip code or searching amazon.com for products, to the more complex tools required by agency staff or travel, task, and community management.

“If it exists on the Web, we can write an interface to it, whether Perl, Java, ASP, whatever,” said Frame. “We’re trying to enable users, so that they don’t have to go to five different places to do their jobs.”

The portal minimizes the need for multiple user ids and passwords to gain access to the various services, while facilitating access to and use of agency resources by diverse populations. Where appropriate to the specific gadget or Web service, users may specify preferences (such as limiting job searches to a specific geographical locale or within a specific category), so that the information retrieved is precisely tuned to parameters appropriate to their needs or wishes.

While some gadgets are “out-of-the-box,” others are specific to administrative procedures and practices required by government agencies. Yet other gadgets are created in order to facilitate interoperability, such as in the instance of a script necessary to permit the email system of the Texas node in the NBII network to work with Lotus Notes. According to Frame, applications being developed by other federal agencies are also being shared via the portal in a “Gadgets and Web Services Showroom.” By working cooperatively with other federal agencies such as the Department of Defense and the National Institutes of Health, NBII continues to keep costs down on infrastructure and development while increasing return on taxpayer dollars by promoting integration of common resources, applications, and Web services. The cooperative development of these interoperable applications from other agencies bolsters federal research and development efforts, as the community is better served by increasing awareness and visibility of the work done across multiple agencies.

As field researchers gather materials and check sources to verify their findings, the need for wireless capabilities is clear. The My.NBII.Gov portal staff is moving rapidly to accommodate those requirements, just as they also provide tools for less-specialized needs of agency personnel, such as flight information feeds for PDAs and phones.

Relevant content is drawn from multiple providers. Both formal and informal communication is enhanced through access to publications, news feeds, bulletin boards, and similar mechanisms of exchange. In conjunction with one private sector content producer, CSA <<http://www.csa.com>>, NBII is sponsoring an e-Forum called "Towards Best Practices" <<http://www.nbii.gov/datainfo/bestpractices/eforum/index.php>>. Users are invited to submit and discuss high quality, science-based publications that define state-of-the-art methodologies, protocols, applications, and analytical tools related to studying and managing biocomplexity. The portal functionality that supports collaborative communities of practice will support a peer-review process of the submissions.

Agency content providers and those associated with the NBII Nodes contribute content through a variety of content management functions. Not only can administrators upload content to the portal for use by various communities, but they also have access to statistical information in order to gauge the most highly accessed content or gadgets for their community.

The content found on the portal resides in a browseable documents directory, consisting of more than 900 different folders that contain a range of file types such as PDF files, Word files, Web documents with embedded hypertext links, even executable software. For the 100,000 documents currently accessible via the portal, each item has what is called a "document card." This "card" contains the metadata for the document, based upon a modified Dublin core standard. For example, opening the document card for a high-resolution vegetation map created by the Oak Ridge National Lab, the user can see the Open Document URL, a description of the map, the specified document, etc. For a species fact sheet, the card may specify appropriate keywords from the indexed document, including the species' scientific name, its place of origin, etc., depending upon the information available in the original document.

Access is facilitated through search capabilities, whether a simple one-word search in a text box on the portal's front page or a more advanced search functionality. The advanced search function enables the user to access both internal resources, such as the NBII Metadata Clearinghouse with its links to datasets, and external resources that are provided by partner organizations such as CSA

and the National Institutes of Health. User authentication controls the appropriate levels of access if content is sensitive in scope or proprietary.

One of several useful applications developed by NBII is the NBII Portal Toolbar, similar in nature to the Google toolbar, but based upon the user's roles and privileges within the NBII community. If a user subscribes to a particular news service, the toolbar can alert the user to new content that may be available. A highlighted button indicates that a message is available, even when the browser is closed or the user is working in a different application. Through the toolbar, the user may navigate directly to one of the portal's communities of practice or access other areas of the portal as necessary.

Collaborative communities exist for specific areas of interest, such as invasive species or geographical information systems, and for specific regional nodes. The communities allow users to manage tasks and projects, discussion threads, scheduling, and other types of collaborative operations that require interaction between users at multiple locations. This innovative "community" functionality allows a university, for example, to offer resources to students taking a specific course, such as Ecology, and it supports video conferencing by guest lecturers as well. This community functionality may be used internally to facilitate the development of proposals, regulations, or software applications, or to review documents within the NBII community. Rather than depend upon email communication with attachments (difficult to track with different versions being seen at various points), members of a particular community use the portal to accumulate and review comments on specific proposals or segments of proposals. Each community may exercise the functionalities available on the portal in ways that best serve their needs and purposes.

In the near future, GBIF and the Inter-American Biodiversity Information Network <<http://www.iabin-us.org>> will be added as communities of practice accessible via the portal.

Recently, the International Council of Scientific Unions <<http://www.icsu.org>> named NBII as a World Data Center for Biodiversity and Terrestrial Ecology. The USGS, the national program office for the NBII, operates this World Data Center, one of over 40 centers worldwide. The NBII will use the portal that has been built as the platform for the international exchange of scientific data with other nations participating in the World Data Center network. The portal will provide the means for researchers to access additional datasets more readily and perform computerized modeling and mapping activities.



Feedback from the question and answer session with regard to the My.Nbii.Gov portal centered on the need for a common agreement within the biodiversity research community as to which system or portal would be the preeminent one. From the perspective of educators, curators and researchers, it would be useful to have a single organization or agency that could be given the authority and on-going responsibility for the creation and maintenance of a portal. The fear is that if multiple sites emerge, all of them incorporating data from multiple repositories such as herbaria, museums and other special collections, then who will be able to ensure the comprehensiveness and inclusiveness of all themes? There seemed to be a general agreement among the *MetaDiversity III* conference attendees that the NBII is the logical choice to be that designated authority, representing the United States with regard to GBIF.

### **Biodiversity Informatics and Education**

Information competency must be part of any discussion regarding large integrated information systems. Presented with a gateway to a wide array of information tools and resources, it is often difficult for students to understand which is the best information resource for their needs, how to retrieve that information and, once retrieved, how to use the information in their educational assignments and tasks.

At *MetaDiversity III*, in discussing the creation of systems and tools in support of biodiversity research goals, **Dr. Ken Klemow, Professor of Biology & Geo Environmental Science, Wilkes University, Wilkes-Barre, PA**, expanded upon the need for education on biodiversity research across all population segments, both professional and amateur. As Bryan Heidorn envisioned the involvement of novice collectors and citizen scientists, and as Tom Moritz emphasized the need for rigorous standardized formats for data, Klemow's presentation highlighted the need for the education of both the providers and users of biodiversity information. It was noted that the provider of data in one environment or context may, at another time or in another context, be the user of information and data.

The overlap of informatics (the application of hardware, software and networks to structured data formatted for easy retrieval) and biodiversity (as represented by genetic, taxonomic, and ecological data) will naturally generate new applications and tools that can support the type of analytical tasks and models required by the biodiversity information community.

As a strong advocate of biodiversity informatics education, Klemow believes that the provision of biodiversity informatics instruction at all educational levels will support the current international, national, and local mandates for heightened awareness of the issues

surrounding allocation of natural resources and biological conservation. In the long term, the effectiveness and productivity of efforts by professionals and local volunteers will be enhanced by an in-depth knowledge of computer applications and data structure when applied to questions of biodiversity involving statistical patterns and analytical models.

As has already been made clear in this report, the biodiversity informatics process begins with the collection of biodiversity data and information, regardless of whether it is recently observed data or a collection of specimens housed by a museum or library. Some level of quality control and quality assurance is imposed on the data, and it is then put into an organized, retrievable format and useful metadata is created. The provider of this biodiversity data then posts it to an online environment, ensuring widespread accessibility to the information.

Interested users can then locate the pooled information, download it, and proceed to perform analysis and/or visualization in order to identify outstanding gaps in the community's knowledge base. The researcher can then perform subsequent tasks of observation and collection to generate new data, thus completing the process and beginning a new cycle. This circular process builds the human storehouse of knowledge.

However, the barriers that inhibit the creation of a fully-integrated biodiversity information system are the same barriers that must be overcome by educators who want to support the work of scientists and others involved in biodiversity research.

Klemow enumerated the barriers regarding the teaching of biodiversity information:

- Issues of scale
- Idiosyncratic data collection methods over time
- Incomplete datasets
- Lack of standardization in both metadata and the datasets themselves
- Legacy data found in paper or physical artifacts, requiring conversion to digital formats
- Lack of a centralized system or repository; balkanization of efforts
- Intellectual and other proprietary rights pertaining to specimens, research results, physical resources, etc.
- Funding

One of Klemow's conclusions was that activities related to biodiversity informatics have focused upon the development of databases and metadata to facilitate the exchange of information. Little effort has been dedicated to education, and yet efforts at education could actually eliminate some of the above-mentioned barriers over time.

For example, the productivity of expert information users would be enhanced if they were fully familiar with the significant repositories of biodiversity data such as those created by the NBII or GBIF. Familiarity with the standards implemented by those organizations would result in better collection methods and more complete information regarding the datasets provided. The productivity of novice collectors and the quality of data gathered in state and localized surveys might be enhanced if students and volunteers were taught the proper use of online taxonomic keys. Facilities, whether academic or museum, would benefit if faculty and staff knew "best practices" for data organization and the generation of standardized metadata.

Thus a biodiversity informatics educational program would:

- Enable individuals to solve specific problems;
- Provide a biodiversity dimension to a broader informatics educational agenda;
- Create expert users of biodiversity information; and
- Create expert providers of biodiversity information.

Klemow briefly sketched the broad parameters of an appropriate biodiversity informatics curriculum but made it clear that strategies for effective education would need to be targeted according to demographics. A K-12 curricula would demand one approach, while adult learners would benefit from another; graduate students would require yet a third educational approach. He noted that in some instances curricula requirements are inflexible, being set by state legislatures. Teachers themselves would need additional support in developing appropriate competencies in the field. There is not a great deal of formalized information as yet regarding biodiversity informatics education.

"Advances in computing power and connectivity allow us to exchange and analyze biodiversity information in ways not possible ten years ago," said Klemow. "However, information exchange can be effectively accomplished only if individuals are educated."

As a follow up to that point, Cynthia Parr of the University of Maryland recommended that appropriate curricula for biodiversity informatics should incorporate those aspects of computing and information retrieval that pertain to the development of the tools

and software applications necessary to achieve the goals and objectives of the biodiversity research community.

The National Science Digital Library was also mentioned as an important complement to the research and education component of biodiversity information.

## REMAINING ISSUES AND WORKING GROUP DISCUSSIONS

Efforts to ensure that the cooperative sharing of information sought by the biodiversity community becomes a reality must overcome obstacles, not only from the technology issues surrounding interoperability, but also from the perspective of the rewards (or lack thereof) associated with data sharing. Bryan Heidorn identified several barriers to the full implementation of an electronic, cooperative information resource. These barriers include intellectual property concerns, attribution for work accomplished, and the long term management costs of electronic systems that would, by virtue of the technological evolution, need to be migrated every 3-5 years in order to maintain their value. Other editorial and archival concerns also need to be addressed, such as the maintenance of authoritative versions of databases, the provision of incentives for "continual authoring," and ensuring ongoing support – editorial and financial – on a national basis.

One initiative attempting to make the cooperative sharing of information a reality is the Biodiversity Commons model, a model presented initially in 2000 by Gladys Cotter and Barbara Bauldock of the USGS/NBII at the International Conference on Very Large Databases held in Cairo, Egypt.<sup>2</sup> As Moritz presented the concept in his 2002 D-Lib paper "Building the Biodiversity Commons,"<sup>3</sup> provision of free, universal access to biodiversity information is a practical imperative for the conservation community.

The Biodiversity Information Commons effort, which draws from a broad constituency of conservation organizations, hopes to create a coherent strategy for addressing the four major constraints on biodiversity information identified previously: current law, cultural norms, current marketplace conditions, and technology. *Zoological Record*, a key information provider and part of the U.S.-based organization BIOSIS, has provided access to their database for the

2 G. A. Cotter and B. T. Bauldock, "Biodiversity Informatics Infrastructure: An Information Commons for the Biodiversity Community," Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000. <<http://www.vldb.org/conf/2000/P701.pdf>>

3 T. Moritz, "Building the Biodiversity Commons," D-Lib Magazine, v.8, n.6, June 2002. <<http://www.dlib.org/dlib/june02/moritz/06moritz.html>>

purpose of analyzing all forms of publications during the past 30 years in order to determine which publishers have contributed most significantly to that database. The group is proposing to negotiate with 5-6 classes of identified publishers in order to develop a common solution for providing free and open access to biodiversity information. Their objective is to go straight to the strategic level in order to break the barriers to free access to content in advance and try and find solutions that will work for all.

### **Addressing Incentives for Sharing**

Discussing ways in which the rewards system can be altered in order to encourage the sharing of biodiversity data, information, and knowledge was a productive activity of one of the conference working groups. The group reported back to the wider audience with substantive ideas. Initially, the group determined that, for their purposes, “reward” was not the correct term to use in the discussion as it has connotations of remuneration. Instead, the group determined that the appropriate term was “incentives.” What incentives might be incorporated into the science of biodiversity that would enable researchers, ecologists and others to encourage the practice of pooling and sharing data and avoid the pitfalls of balkanized collections of datasets, research material, and other biodiversity content?

*What incentives might be incorporated into the science of biodiversity that would enable researchers, ecologists and others to encourage the practice of pooling and sharing data, and avoid the pitfalls of balkanized collections of datasets, research material and other biodiversity content?*

The group listed five separate levels at which the incentives issue might be addressed:

- At the individual level – the scientists, researchers and conservationists who are primarily responsible for generating these datasets, research materials and other sources of biodiversity information;
- At the professional level (the cultural norms of Science);
- At the institutional level – museums, non-governmental organizations, and universities;
- At the national (governmental) level; and
- At the commercial level.

The group was careful to stipulate that some limits should be set on shared access, based upon the sensitivities within specific sectors and at various sites.

### **INDIVIDUAL SCIENTISTS, RESEARCHERS, AND CONSERVATIONISTS**

When dealing with individual scientists, researchers, and conservationists, it is important to recognize the importance of the “publish or perish” strictures related to promotion and tenure, grant applications, and the prestige factor.

Publication demonstrates the usefulness and utility of the work performed by the individual scholar. Datasets, databases, and analytical software tools represent some of the materials from which numerous papers can be developed by those working in biodiversity. Incentives, therefore, must provide accommodation for the preservation and integrity of such materials as well as a shift in the recognition factor so that the creation of a useful database or software tool may actually be considered a “publication.”

Creators of datasets are often concerned about protecting the integrity of their work. Internationally, such as in Canadian and Mexican copyright law, such concern is referred to as the “moral rights of authorship,” the right to be identified as the author of a work and the right not to have it subjected to derogatory treatment by prejudicial additions or alterations. The need for protection must be considered a legitimate concern in complex environments where, as illustrated by such projects as DiscoverLife.org, portals act as aggregators of content. Respecting the need for attribution and maintaining data integrity within the limits of practice is very important. Indeed, with regard to datasets, it is essential to protect the original integrity of the data. It should be possible to go back to the original source in as many instances as possible in order to scrutinize the original data.

The working group believed that there were additional ways in which the sharing ethic could be made more attractive to the working individual. One recommendation was the use of more sensitive measures of impact, using Web auditing tools. Said Tom Moritz, AMNH and spokesperson for this working group, “One of the incentives that we can provide to people who provide these sources of [biodiversity] information is to give them back more sensitive reflections of how and who is using their data within the limits of confidentiality and privacy.”

Peer review was also identified as an area in which the cooperative sharing of data could be promoted. The working group suggested that it might be useful to develop more conventional forums, either via journals or other online sources, in which peer review could occur. Such forums could be used in the conventional way for purposes of promotion or incentives to individuals to generate and share datasets and other forms of content.

It was also suggested that there might be new forms of “surrogate peer-review,” digital activities that are outside the realm of conventional peer-review processes, but that are well-suited to the digital environment. One example given was the National Library of Medicine’s practice on the PubMed Web site to permit the posting of original material that had not been formally peer-reviewed, but that had been “vouched for” or sponsored by at least two eligible researchers

<<http://www.nih.gov/about/director/pubmedcentral/pubmedcentral.htm>>. Other modifications or adaptations of the traditional peer review model may also be possible in the Web environment.

#### **THE PROFESSIONAL LEVEL: THE CULTURE OF SCIENCE**

A requirement currently exists within the culture of science for attribution and citation in order to prevent or minimize plagiarism. Therefore, both appropriate citation practices and penalties for plagiarism are established cultural norms. The question is whether these norms can be extended in a logical way in order to require full citation and attribution in the digital realm.

Data sharing *could* be made a cultural requirement. Both the National Aeronautics and Space Administration and the National Oceanic and Atmospheric Administration have sunset regulations, legal provisions for the diffusion of government-funded research, addressing sharing of data resulting from research funded by these agencies.

As Moritz commented, “They don’t send the data sharing cops out to chase you down if you haven’t done it. It’s not beyond the reach of our community to say that, within the area of biodiversity, such requirements might exist for the sharing of data and information.”

He referenced the example of the GenBank model <<http://www.ncbi.nlm.nih.gov/Genbank/index.html>>, noting data sharing as a normative agreement among the biotech community in pursuit of a common database. If a member of that community were to publish data, it has to be deposited in GenBank as a requirement of publication. GenBank is a recent example of a community that has established data sharing as a requirement.

#### **INSTITUTIONAL INCENTIVES (MUSEUMS, NON-GOVERNMENT ORGANIZATIONS, AND UNIVERSITIES)**

Incentives at the institutional level must be approached with an understanding that there are very real proprietary concerns associated with intellectual property. Technology transfer officers, general counsels, and other appropriate staff have an obligation to protect and manage the intellectual assets of their institutions. Consequently, such individuals will err on the conservative side when

confronted by external efforts and pressure to share or provide open access to their data or information. Underlying their sense of obligation is a fear of loss in revenues or in competitive stature with other institutions. One counter to such concerns might be for museum and university accreditations to include appropriate benchmarks with regard to data sharing.

Another recommendation put forward by Richard Huber, Principal Environmental Specialist, Organization of American States, was that professional societies or networks could perform external evaluations of institutional or organizational Web sites with regard to their willingness to share data. The Organization of American States has performed this type of evaluation of Web sites in support of the Inter-American Biodiversity Information Network. It was noted in the working group discussions that some of the most prominent conservation organizations did not rate very highly in such an evaluation; it was also noted, however, that this exercise was a fairly subjective process. Perhaps a set of more rigorous standards for site evaluation could be developed with the objective of providing institutional incentives to share data and information. If a generally-accepted external rating system resulted in an institution being judged a “data hoarder,” it might act as an incentive for that institution to think more about sharing and making their data freely available.

#### **GOVERNMENTAL LEVEL**

Governmental concerns are related to the potential exploitation of national patrimony. Fears of bioprospecting by developed nations in areas of the world still rich in biodiversity, for example, fuel a cautious response by many developing and transitioning countries. The Convention on Biological Diversity <<http://www.biodiv.org>> requires repatriation of information, and that requirement was established for a reason. There is a perception that many institutions, particularly institutions from highly developed nations, have been guilty in the past of appropriating genetic resources, a form of “biopiracy.” Many of the concerns that have been mentioned with regard to individuals and institutions are applicable to governments as well. If governments can be assured that data is not going to be taken and used for commercial purposes – bioprospecting – and if models can be developed that clearly protect against such abuse of local resources, then such models will alleviate this concern, at least in part. However, they will not entirely solve the problems associated with biopiracy, and bioprospecting (see “*Bioprospecting Has Failed: What Next?*” at <<http://www.grain.org/seedling/seed-02-10-7-en.cfm>> for a general explanation of how the promise of bioprospecting has disappointed proponents of the practice).

The emerging Biodiversity Information Commons model referenced earlier suggests how an information-sharing regime can be constructed to protect national interests (for example, to protect against resource exploitation, bioprospecting, etc.) and to insure that essential biodiversity data, information, and knowledge can be shared.

### **INCENTIVES FOR THE COMMERCIAL SECTOR**

The working group was unable to discuss in great detail the impact of information sharing by commercial organizations, but did provide a model that might be considered. The World Health Organization Health InterNetwork Access to Research Information (WHO-HINARI) Project <<http://www.healthinternetwork.org>> has negotiated with commercial publishers to provide free access to their literature – the digitized literature of more than 2,100 journals – from a group of 69 countries that have a per capita gross national product of less than \$1,000. Said Tom Moritz, “That’s not a trivial thing. Elsevier is, in many people’s minds, the demon of the commercial publishing realm, and yet they have agreed to participate in this WHO-HINARI project. And the BioOne system is now, I believe as of today, also participating in that. So there are approaches that can be made to commercial interests to encourage data sharing, although they may not entirely solve the problem.”

### **Information Discovery**

One participant offered, “I work for an institution that includes the phrase “information-sharing” as part of its mission statement. And we interpreted and implemented the phrase in such a way that it meant not just having it shareable, but also discoverable. A library has books that are shareable, but without a card catalog or an index, they’re not discoverable. I was surprised to see how many agencies didn’t have that explicitly in their mission statement. I don’t know if that’s specifically government, non-governmental organizations, non-profit, library whatever, but I think that is something that has to do with incentives. Examine your charter, examine your fundamental purpose, and see if sharing of data and rendering your information discoverable is part of it, because sometimes that’s what you need to have in order to justify any investment in doing that.”

Moritz responded that his working group had two discussions that might be relevant to that concern. First is what is referred to as the “deep Web” problem. Much of the material on the Web is buried in databases. At the AMNH digital library Web site <<http://www.amnh.org>>, some 20,000 pages of scientific material are up on the Web as part of one Web site alone. But this information is not accessible directly at the Web site; the user has to run a search to be able to pull results from those pages. A Web crawler would not

have been able to index the scientific material directly. This is a generic problem. (AMNH, added Moritz, may have addressed this problem in the new version of their digital library.)

Over each of the past 15 years, the American Museum of Natural History has had a significant annual revenue stream of more than \$100,000 from licensing images from its institutional archives and special collections. “But,” said Moritz, “we’ve been having a very active discussion about this question of what is consistent with our mission. As much as many of us want to just keep surviving and keep funding our institutions, revenue is not our mission fundamentally. Our mission is really about educating people about the natural world, so the problem became: how can we actually discriminate between those sources of revenue that were appropriate? In our case, we figured those that were purely commercial uses of our images – advertising for soap or microwave ovens – were legitimate; we should get full market value for the use of our images. On the other hand, if a non-profit comes to us and says we want to run a conservation campaign on biodiversity in upstate New York, our view at AMNH is that we should offer everything for free. We should just see that as our mission.” An example of commercial textbook publishers whose work is mission-related but still commercial is considered by AMNH to be a “middle ground.” In such cases, AMNH would offer their images at cost. “But we *have* been thinking through exactly those sets of things in terms of the mission of the institution,” concluded Moritz, “as opposed to just that reflexive thing of ‘however we can get revenue, we ought to be pulling it in the door’ and I think it is important for all of us to be thinking along those lines as well.”

### **Repatriation Issues**

Gary Rosenberg of the National Academy of Science raised another concern of the biodiversity and museum communities, that of the repatriation of specimens. One concern of institutions is that, as more of this data becomes available, there will be demands from some countries to return specimens if there’s no evidence that they were collected legally. Rosenberg asked if there was some way agreements could be put in place wherein, in return for making data available, for anything collected before the year 2000 the circumstances of collection will be ignored. Rosenberg offer that such agreements have been concluded in the past. “I think that the Association of Systematic Collections did negotiate an agreement with the Parks Service that things, sitting in museum collections, collected on Park Service land can stay in the museums collections. And they’re not going to go through the battle about ownership.”

Moritz responded, “The Convention on Biological Diversity specifies repatriation of information, and I think that’s really a critical thing. Repatriating objects (such as the Elgin Marbles) is a problem.”

John Pickering of DiscoverLife.org took a slightly different view of the issue. “I’d counter Gary [Rosenberg]’s statement because if we really do our job well why would you want the specimen? It’s on the Web. Cost of physical storage is a considerable expense. If you’ve got access to the digital images there, you’ve run it through and put it in GenBank, what else do you require?”

### **Standards Revisited**

Why consider standards for the dissemination of biodiversity data? Are standards a requirement? The working group dealing with this topic was able to identify several examples in which the implementation of data standards would improve the access to and use of biodiversity data. Sue Thompson of the Pennsylvania BioDiversity Partnership acted as spokesperson for the group.

The use of standards is appropriate when:

- Collectors of data want to share that data with external audiences. There is expediency to standards when there is a need to communicate common concepts;
- Researchers are interested in sustaining the long-term use of data-set collections for analytical purposes in diagnosing issues of invasive species or disappearing species. Standards ensure long-term viability of archived data; and
- The creator of a dataset is initiating a new project requiring the collection of a novel dataset or the use of a new technology. Under such circumstances, it is ideal to follow “best practices” and use a form of community or accepted standard. This ensures quality assurance and quality control. There is a value added to a standardized dataset, simply by virtue of the ability to share and combine that data with existing data in exploring new concepts and concerns.

In circumstances where the researcher is using off-the-shelf software, whether consumer software such as Windows or an application specific to biodiversity, use of that software represents a *de facto* standard in the structuring of data. Technology forces compliance with existing standards. For purposes of cross-walking and mapping, use of standards and technology will enable consistency and interoperability, even though on a limited level, it may appear that the use of such standards will make datasets non-compliant.

Whether following external standards or not, an internal consistency needs to be applied to data. Standards, especially in the biological world, need to accommodate legacy data and incomplete data. Essentially, most of the data on which biodiversity and conservation decisions are based are rooted in legacy collections that are housed in museums. Such collections were usually assembled well before the advent of standards, and thus researchers today must be able to accommodate the gaps and inaccuracies frequently to be found in such specimen collections.

Complaints about standards wasting time and effort frequently arise when researchers and scientists apply a standard without taking a long-term view of their project. The use of standards depends upon the goal. What is the objective of the project? Standards are not identical for all types of studies or projects. The use of the standard depends upon the study involved. Is conservation the primary focus? Will the data be shared? In many cases, global datasets in biodiversity are discussed without taking into consideration the true scope of the community who will ultimately make use of that data. Does the global community need to be involved if the conservation decision is for Napa Valley, California? It is important to consider just what the end product of the study or project will be. Standards should not be used just for the sake of using them. There must be a specific purpose to their utilization.

### **Who Should Set Standards?**

Data providers and end users should be included in discussions on standards. There are multiple groups working on standards (see <<http://www.convergedigest.com/StandardsBodies.htm>>, including the Taxonomic Data Working Group <<http://www.tdwg.org>>).

In the technology arena, technology vendors bear a special burden because they establish *de facto* standards through market share. The larger the market share held by a vendor, the more that vendor’s standards are accepted. Vendors need to be cautious and need to communicate frequently with the user community in developing their product standards. Frequently, *de facto* standards set by market share do not serve the needs of the user community. One of the models discussed by the working group was the Open GIS Consortium <<http://www.opengis.org>> that began with vendor funding and which has, over time, developed good standards for geographic information systems. The Open GIS Consortium is a growing organization, one that includes OBIS, NBII, academic groups, and others. It is one type of model to consider.

The group questioned whether governments should set standards, given that governments are large producers of data as well as intensive users of data. Bruce Westcott’s presentation had touched

on the fact that the U.S. government had been trying to move away from being closely involved with the development of federal standards, moving instead towards compliance with international standards groups. The biodiversity community sees this as a necessity in promoting the *global* interoperability of data.

### **Problems with Standards**

Standards are facultative; if they are not accepted, they become dysfunctional. With regard to biodiversity, there are really no accepted community standards. Some common fields across databases provided by museums and other organizations (fields such as locations, nomenclature, genus, species, scientific terminology, etc.) do exist, and there have been some limited attempts to standardize those data elements. The Darwin Core elements, developed by Dave Vieglais of the University of Kansas, represents one such attempt. It is now evolving into Version 2.0., and is still in development.

Sue Thompson, the working group's spokesperson, did note one problematic aspect with regard to primary data and derived data. It is important to capture the primary data in a fielded structure. A lower priority can be given to the derived data, as those fields can be developed in a later timeframe. In addition, an issue related to legacy data involves adding a non-existent precision to the information. There is a tendency to want to impose precision to information as standards are applied, especially with regard to collections that were not precisely geo-referenced when originally gathered. The community needs to be able to enhance such legacy data imprecision without skewing the data's accuracy and value.

John Pickering, DiscoverLife.org, offered that he was not a big fan of standards but that he had one recommendation for everyone: "Whatever you do, document it and get it on the Web. So if you've got an XML schema, put it on the Web, and then people like me will write translators and not worry about what standard you're using. I assume you're using something and I want to know what you're doing. And that's my recommendation: DOCUMENT IT."

Concerning the necessity of documentation, Gary Rosenberg, Academy of the Natural Sciences, added, "Document methods across the board (including methods of collection and samplings so people can do the gap analysis). We need to know not just the positive aspects of your data, but also the negative aspects – enough so that others within the community can understand fully what the data may be used for!"

### **END NOTE**

The *MetaDiversity III* conference represents a milestone, marking the progress of biodiversity and ecosystem informatics towards the desired goal of a global information system. Clearly progress has been made on the recommendations that emerged from prior MetaDiversity conferences, but advancement needs to be made in the areas identified above in order to achieve the type of information resources required for intelligent decision-making and for the establishment of priorities in protecting our world's biodiversity.

**APPENDICES****Appendix A: Program****Metadiversity III:****Global Access for Biodiversity Through Integrated Systems****March 31 - April 1, 2003**

*MetaDiversity III* was held under Cooperative Agreement 02HQAGO111 between the NBII/U.S. Geological Survey & NFAIS

**Final Program****Monday, March 31****8:15am - 9:00am** Registration and Continental Breakfast**9:00am - 9:30am** Welcome and Establishing The Context

Results from MetaDiversity II will be presented with an overview of anticipated outcomes of MetaDiversity 3

**9:30am - 10:15am** Keynote Address

Meredith Lane, Global Biodiversity Information Facility

**10:15am - 10:30am** Break**10:30am - 11:30pm** Information Technology and Information Professionals

P. Bryan Heidorn, Graduate School of Library and Information Science, University of Illinois, Urbana Champaign

Tom Moritz, Boeschstein Director, Library Services, American Museum of Natural History

**11:30am - 12:00pm** Questions and Discussion**11:45am - 12:15pm** Biodiversity Informatics and Education

Ken Klemow, Professor of Biology & GeoEnvironmental Science, Wilkes University

**12:15pm - 1:10pm** Luncheon**1:15pm - 2:00pm** Case Study in Field Research

From the Tropical Forest Photographer to Global User: Simplifying the Journey from Digitals to Comparative Web Pages for Massive Numbers of Images

Daniel H. Janzen, Professor of Biology and Thomas G. and Louise E. DiMaura Term Chair, University of Pennsylvania

**2:00pm - 3:00pm** International Activities in Biodiversity Systems

Steve Clemants, Vice President of Science, Brooklyn Botanical Garden

Gary Rosenberg, Associate Curator, Academy of Natural Sciences

John Pickering, Associate Professor, Institute of Ecology, University of Georgia

**3:00pm - 3:15pm** Peer Review Commentary**3:15pm - 3:30pm** Break**3:30pm - 5:00pm** The NBII Portal Experience

Michael T. Frame, NBII Research & Technology Director, USGS

**5:30pm - 7:00pm** Reception**Tuesday, April 1****8:30am - 9:00am** Continental Breakfast**9:00am - 9:45pm** Keynote Presentation

Eric Hellman, President, Openly Informatics

**9:45am - 10:00am** Break**10:00am - 11:30am** The Role of MetaData and Standards

Jill Trubey, Metadata Coordinator, Florida Fish & Wildlife Commission, Florida Marine Research Institute

Bruce Westcott, Spatial Metadata Consulting

Larry Sugarbaker, Vice President & Chief Information Officer, NatureServe

**11:30am - 12:00pm** Working Group Deliberations**12:00pm - 1:30pm** Working Groups Working Luncheon**1:30pm - 3:00pm** Working Groups Report



## Appendix B: List of Attendees

### Participants

#### Bill Barnett

Vice President and Chief Information Officer

#### The Field Museum

wbarnett@fieldmuseum.org

#### Marciela Canepa-Montalvo

USDE Consultant 08

#### Organization of American States

usdeintlo@oas.org

#### Marta Ceroni

Associate Professor

#### University of Vermont

marta.ceroni@uvm.edu

#### Steven Clemants

Vice President of Science

#### Brooklyn Botanic Garden

steveclemants@bbg.org

#### Nancy Cothran

Master of Forest Science Candidate,  
School of Forestry and Environmental Studies

#### Yale University

nancy.cothran@yale.edu

#### Christopher Dunn

Director of Research

#### The Morton Arboretum

cdunn@mortonarb.org

#### Alvaro Espinel

Manager, Database Operations

#### Center for Applied Biodiversity Science

#### Conservation International

a.espinel@conservation.org

#### Michael T. Frame

NBII Research & Technology Director

#### USGS

mike\_frame@usgs.gov

#### Andrea Grosse

Biodiversity Information Specialist

#### USGS-NBII

agrosse@usgs.gov

#### P. Bryan Heidorn

Assistant Professor

#### University of Illinois

pheidorn@uiuc.edu

#### Eric Hellman

President

#### Openly Informatics

eric@openly.com

#### Richard Huber

Principal Environmental Scientist

#### Organization of American States

RHuber@oas.org

#### Mariana Ibarcena-Escudero

USDE Consultant 08

#### Organization of American States

usdecpr@oas.org

#### Daniel H. Janzen

Professor of Biology

#### University of Pennsylvania

djanzen@sas.upenn.edu

#### Dr. Paul Kanciruk

Senior Research Staff

#### Oak Ridge National Laboratory/DOE

pkk@ornl.gov

#### Richard T. Kaser

Vice President, Content

#### Information Today

kaser@infoday.com

#### Bruce Kiesel

Director, Knowledge Base Development

#### BIOSIS

bhkiesel@mail.biosis.org

#### Kenneth M. Klemow

Professor of Biology and

GeoEnvironmental Science

#### Wilkes University

kklemow@wilkes.edu

#### Meredith A. Lane, Ph.D.

Communications Officer, GBIF

Secretariat

#### Global Biodiversity Information Facility

mlane@GBIF.org

#### Bonnie Lawlor

Executive Director

#### NFAIS

blawlor@nfais.org

#### James Lester

Director, Environmental Group

#### Houston Advanced Research Center (HARC)

transome@harc.edu

#### Joanna McCaffery

Collections Database Architect

#### The Field Museum

jmccaffrey@fieldmuseum.org

#### Joan Meranze

Manager, Journals

#### ASME International

meranzej@asme.org

#### Tom Moritz

Boeschstein Director, Library Services

#### American Museum of Natural History

tmoritz@amnh.org

#### Paul Morris

Biodiversity Information Manager

#### Academy of Natural Sciences

mole@morris.net

#### M.P. Mulligan

Physical Scientist

#### USGS/BRD/CBI

mike\_mulligan@usgs.gov

#### Celia Najara-Dinicola

Web Manager/NBII Liaison

#### Nature Serve

celia\_dinicola@natureserve.org

#### Jill O'Neill

Director, Planning & Communication

#### NFAIS

jilloneill@nfais.org

#### Cynthia Parr

Assistant Research Scientist

#### University of Maryland

csparr@umd.edu

#### John Pickering

Associate Professor

#### University of Georgia

pick@discoverlife.org

#### Robert Pritchett

Product Analyst

#### BIOSIS

rpritchett@biosis.org

#### Arturo Restrepo

USDE Consultant 08

#### Organization of American States

ARestrepo@oas.org

#### Gary Rosenberg

Associate Curator

#### Academy of Natural Sciences

Rosenberg@ansp.org

#### Stephen Sand

Lead Editor

#### BIOSIS

ssand@mail.biosis.org

#### Annie Simpson

Invasive Species Theme Coordinator

#### U.S. Geological Survey

asimpson@usgs.gov

#### Larry Sugarbaker

Vice President and Chief Information Officer

#### NatureServe

larry\_sugarbaker@natureserve.org

#### Sue Thompson

President

#### Pennsylvania Biodiversity Partnership

Thompson@pabiodiversity.org

#### Jill Trubey

MetaData Coordinator

#### Florida Marine Research Institute

jill.trubey@fwc.state.fl.us

#### Ferdinando Villa

Research Professor

#### University of Vermont

fvilla@uvm.edu

#### Bruce Westcott

Consultant

bspatal@together.net

**Appendix C: List of Acronyms**

<b>AMNH</b>	American Museum of Natural History
<b>BIBE</b>	Biological Information Browsing Environment
<b>CAMRA</b>	Coastal and Marine Resource Assessment program
<b>CONABIO</b>	National Commission for the Knowledge and Use of Biodiversity (Mexico)
<b>FGDC</b>	Federal Geographic Data Committee
<b>FMRI</b>	Florida Marine Research Institute
<b>GBIF</b>	Global Biodiversity Information Facility
<b>GIS</b>	Geographic Information System
<b>ISO</b>	International Organization for Standardization
<b>MARC</b>	Machine-Readable Cataloging
<b>NBII</b>	National Biological Information Infrastructure
<b>NISO</b>	National Information Standards Organization
<b>NYMF</b>	The New York Metropolitan Flora project
<b>OBIS</b>	Ocean Biogeographic Information System
<b>USGS</b>	U.S. Geological Survey

**Appendix D: Index of Important Organizations, Key Words, and Links to Related information on the Web\*****Academy of Natural Science Philadelphia:**<http://www.acnatsci.org/>**American Museum of Natural History (AMNH):**<http://www.amnh.org/>**American Museum Congo Expedition 1909 – 1915:**<http://diglib1.amnh.org/description.html>**American National Standards Institute (ANSI):** <http://www.ansi.org/>**Area de Conservacion Guanacoste (ACG):** <http://janzen.sas.upenn.edu/>**Berners-Lee, Tim:**<http://www.bioone.org/bioone/?request=index-html>**Best Practices:** <http://www.nbii.gov/datainfo/bestpractices/eforum/index.php>**Biodiversity:** <http://www.defenders.org/bio-bi03.html> &<http://www.wri.org/biodiv/bri-ntro.html>**Biological Information Browsing Environment (BIBE):**<http://www.biobrowser.org/>**BioOne:** <http://www.bioone.org/bioone/?request=index-html>**BIOSIS:** <http://www.biosis.org/>**Bioprospecting:** <http://www.nature.nps.gov/benefitssharing/whatis.htm>**Biopiracy:** <http://webpages.charter.net/westons/biopiracy.html>**Cambridge Scientific Abstracts (CSA):** <http://www.csa.com/>**Conabio:** <http://www.conabio.gob.mx/>**Convention on Biological Diversity(CBD):** <http://www.biodiv.org/default.aspx>**Darwin Core Standard:**<http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV1>**Discover Life:** <http://www.discoverlife.org/>**Dublin Core Metadata Standard:** <http://dublincore.org/>**Federal Geospatial Data Center (FGDC):** <http://www.fgdc.gov/>**Florida Marine Research Institute (FMRI):** <http://www.fmri.usf.edu/>**GenBank:** <http://www.psc.edu/general/software/packages/genbank/genbank.html>**Global Biodiversity Information Facility (GBIF):** <http://www.gbif.org/>**Global Positioning System (GPS):** <http://www.trimble.com/gps/index.html>**Global Species Databases (GSDs):**<http://www.indexfungorum.org/GSD/GSD.htm>**Health InterNetwork Access to Research Information (HINARI):**<http://www.healthinternetwork.org/>**Integrated Taxonomic Information System (ITIS):**<http://www.itis.usda.gov/>**Inter-American Biodiversity Information Network (IABIN):**<http://www.iabin-us.org/>**International Council of Scientific Unions (ICSU):** <http://www.icsu.org/>**International Organization for Standardization (ISO):**<http://www.iso.ch/iso/en/ISOOnline.frontpage>

**MARC Standards:** <http://loc.gov/marc/>

**Metadata:** <http://walrus.wr.usgs.gov/infobank/programs/html/definition/meta.html>

**MY.NBII.Gov Portal:** <http://my.nbii.gov/>

**National Biological Information Infrastructure (NBII):**

<http://www.nbii.gov/>

**National Commission for the Knowledge and Use of Biodiversity**

**(Conabio):** <http://www.conabio.gob.mx/>

**National Geospatial Data Clearinghouse:** <http://clearinghouse1.fgdc.gov/>

**National Information Standards Organization (NISO):**

<http://www.niso.org/>

**National Science Collection Alliance (NSCA):** <http://www.nscalliance.org/>

**NatureServe:** <http://www.natureserve.org/>

**National Spatial Data Infrastructure (NSDI):**

<http://www.fgdc.gov/nsdi/nsdi.html>

**New York Metropolitan Flora Project (NYMF):** <http://www.bbg.org/sci/nymf/>

**NFAIS (formerly the National Federation of Abstracting and**

**Information Services):**

<http://www.nfais.org/>

**Ocean BioGeographic Information System (OBIS):**

<http://www.iobis.org/OBISPortal/> & <http://data.acnatsci.org/obis/>

**Open Geographical Information Systems (Open GIS):**

<http://www.opengis.org/>

**OpenKey:** <http://www.isrl.uiuc.edu/~openkey/> and <http://www.ibiblio.org/openkey/>

**OpenUrl Standard:** <http://www.niso.org/standards/resources/OpenURL-release.html>

**Plumtree:** <http://www.plumtree.com/>

**PolyClave:** <http://www.library.utoronto.ca/polyclave/>

**Resource Description Framework (RDF):** <http://www.fgdc.gov/nsdi/nsdi.html>

**Semantic Web:** <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>

**Spatial Metadata Management System (SMMS):**

<http://imgs.intergraph.com/smms/>

**Tablet PCs:**

[http://searchwin2000.techtarget.com/gDefinition/0,294236,sid1\\_gci509982,00.html](http://searchwin2000.techtarget.com/gDefinition/0,294236,sid1_gci509982,00.html)

**Telenature:** <http://www.isrl.uiuc.edu/~telenature/>

**U.S. Geological Survey (USGS):** <http://www.usgs.gov>

**World Resources Institute:** <http://www.wri.org/>

**Zoological Record:** [http://www.biosis.org/products\\_services/zoorecord.html](http://www.biosis.org/products_services/zoorecord.html)

## Additional Reading:

**MetaDiversity I:**

[http://www.nfais.org/publications/metadiversity\\_preprints\\_contents.htm](http://www.nfais.org/publications/metadiversity_preprints_contents.htm)

**Metadiversity II:** <http://www.nfais.org/publications/metadiversityII.pdf>

**Berners-Lee, "The Semantic Web", Scientific American, April 2002.**

<http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>

**Cotter, Gladys A., Bauldock, Barbara T., Biodiversity Informatics Infrastructure: An Information Commons for the Biodiversity Community,** <http://www.vldb.org/conf/2000/P701.pdf>

**Moritz, Thomas, "Building the Biodiversity Commons", D-Lib Magazine, June 2002,**

<http://www.dlib.org/dlib/june02/moritz/06moritz.html>

**Rodriguez, Sylvia, "Bioprospecting has Failed - What Next?"**

<http://www.grain.org/seedling/seed-02-10-7-en.cfm>