The 6th International Conference on Information Technology and Applications (ICITA 2009)

National Biological Information Infrastructure: Accessing Data on a Country's Living Capital

Thomas Hermann, Gladys Cotter, Tom Lahr, and John Hill

Abstract— **Biological resources, from the genetic information within organisms to the life-sustaining services provided by ecosystems, are vital to human well-being, including the economy [1]. In order for researchers, natural resource managers, decision makers, and citizens to understand these resources and maintain them, it is important that the capacity exists for accessing, sharing, integrating, and utilizing pertinent data and information. The mission of the U. S. Geological Survey (USGS) Biological Informatics Office Program is to create the informatics framework (through development and implementation of the National Biological Information Infrastructure [NBII]) to provide the scientific content and develop the pubic and private partnerships needed for understanding and stewardship of the Nation's biological resources. This paper describes the goals, components, associated bioinformatics technologies, services, partnerships, scientific and operational applications, and future direction of the NBII.**

Index Terms— **Biology, Bioinformatics, Biological Informatics, Ecology**

I. INTRODUCTION

It is now generally accepted that improving and conserving the environment is compatible with growing a Nation's economy [2] and well-being of its citizens. Therefore, there is a need for an extensive and frequently updated knowledge base in order to optimize the union between the environment and economy.

By harnessing advances in information theory and large capacity computational systems, one can better understand and manage living resources. In addition, there is the need to address constantly emerging environmental issues (e.g., climate change, invasive species, decline in the number of species [i.e., amphibians, pollinators], zoonotic [human/wildlife] diseases, and renewable energy) that annually have impacts in the billions of dollars. Therefore, the National Biological Information Infrastructure (NBII) was created with the mission acquiring and integrating the nation's biological and ecological data and making it accessible to the scientific and conservation management community. The NBII catalogs and links to data and information resources, provides searches of Web sites, develops tools for the integration, mining, and display of data, and synthesizes the latest information on selected topics [3]. It also actively supports the development and use of standards and protocols.

This work was supported in part by the U.S. Geological Survey (USGS), Biological Informatics Office and its partner organizations.

Thomas A. Hermann is with the U.S. Geological Survey, Biological Informatics Office, 302 National Center Reston, VA 20192 USA (telephone: 703-648-4211, e-mail: thomas_hermann@usgs.gov).

Gladys Cotter is with the U.S. Geological Survey, Biological Informatics Office, 302 National Center Reston, VA 20192 USA (telephone: 703-648-4182), e-mail: gladys_cotter@usgs.gov).

Tom Lahr is with the U.S. Geological Survey, Biological Informatics Office, 302 National Center Reston, VA 20192 USA (telephone: 703-648-4222), e-mail: tom_larh@usgs.gov).

John M. Hill is with the ICSU World Data Center for Biodiversity and Ecology, 70 Indian Clover Drive, The Woodlands, TX USA (telephone: 713-828-8389, e-mail: jhilltx@gmail.com.

II. BACKGROUND

The NBII was created primarily through the recommendations of a National Research Council (NRC) in 1993 [4]. The NRC recognized the need for a broad, collaborative program involving all sectors to provide access to data and information on the Nation's biological resources. NRC realized the need to establish mechanisms for the efficient, coordinated collection and dissemination of new data and information, as well as the tools for integration and analysis. The NBII also assists in maximizing the nation's annual public investment in biological research and data collection, estimated at more that $500 million [5]. This helps avoid duplicative data collection and makes for discoveries possible, through access to information. The NBII also serves as a more mature model of E-government. E-government emphasizes the use of enhanced technology, including the Internet, to make it easy for citizens and business to interact with government, save taxpayer dollars, and simplify the business-to-government transactions. NBII also typifies many E-government emerging trends:

- End-user focus,
- Defined and scalable milestones,
- Public-Private partnerships,
- Alliances with stakeholders, and
- Interagency cooperation.

As a distributed system, the NBII links and leverages information sources in biology, ecology, geography, and environmental science across the nation and the world. NBII and its partner organizations also develop new standards, tools and technologies for NBII customers to find, integrate, and apply biological resources information.

NBII is being developed as a node based structure to ensure partnerships and information from all sectors of society [5]. Nodes are interconnected NBII entry points, which in-turn feed the requirements of a national system. NBII nodes, represent the

following functions:

- Regional - which have a geographic orientation. This allows local data and associated regional biological issues to be addressed.
- Thematic - which focus on a specific biological issue that expands across regional boundaries at the national and international levels. Example themes include: bird conservation, invasive species, fisheries and aquatic resources, and wildlife disease-health.
- Infrastructure and Knowledge Management - which is responsible for data standards, appearance and tools, as well as data dissemination.

Based on partnerships, interagency cooperation exists between the NBII and such federal agencies as: The National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), Environmental Protection Agency (EPA), and National Oceanic and Atmospheric Administration (NOAA). Non-profit partnerships include such organizations as the Ecological Society of America (ESA), Conservation International (CI), National Science Collections Alliance, Association of Fish and Wildlife Agencies, Birdlife International, and NatureServe, as well as numerous universities.

The NBII also participates in a wide range of global partnerships designed to support the exchange and use of biodiversity information across national and regional borders. Global initiatives operate on a worldwide level, helping integrate and utilize both national-level biodiversity information and regional networks. These initiatives provide the information required by countries to improve decision-making, particularly for issues at the interface of human development and biodiversity conservation. Many of these initiatives are also interoperable with the national and regional frameworks with which the NBII engages. Examples of NBII's global partners and initiatives include, but are not limited to:

- Global Biodiversity Information Facility (GBIF) - which has a focus of species and specimen data. NBII serves as the U.S. node for GBIF (http://www.gbif.org).
- Clearing-House Mechanism (CHM) of the Convention on Biodiversity (CBD) - which is designed to facilitate worldwide scientific cooperation and information on biodiversity data. (http://www.cbd.int/chm)
- Inter-American Biodiversity Network (IABIN) - an initiative of the Summit of the Americas, which extends the principles of standards and cooperation in the management of biological information at the hemispheric scale. NBII serves as the U.S. node for IABIN. (http://www.iabin.net)
- Global Invasive Species Information Network (GISIN) – which was formed to provide a platform for sharing invasive species information at a global level, via the Internet and other digital means. (http://www.gisinetwork.org)
- International Council for Science (ICSU) World Data Center System – since data constitute the raw material of scientific understanding, the World Data Center System works to guarantee access to solar, geophysical and related environmental data. It serves the whole scientific community by assembling, scrutinizing, organizing and disseminating data and information. The NBII is the U.S. host for the World Data Center for Biodiversity and Ecology (WDC-BE). The NBII and WDC-BE are also the U.S. hosts for the new World Data Center for Biodiversity and Human Health (the first WDC in all of Africa). (http://www.ngdc.noaa.goc/wdc and/or http://www.ngdc.noaa.gov/**wdc**/guide/**wdc**guide.html)
- Consortium for the Barcode of Life (CBOL) - is an international initiative devoted to developing DNA barcoding as a global standard for the identification of biological species. DNA barcoding is a new technique that uses a short DNA sequence from a standardized and agreed-upon position in the genome as a molecular diagnostic for species-level identification. DNA barcoding enables easy and cost-effective species identification. (http;//www. barcoding.si.edu)
- Global Forest Information Service (GFIS) - provides the framework to share forest-related data and information through a single gateway. It promotes the dissemination and sharing of forest and tree-related information and knowledge among the global forestry community by developing common information exchange standards, building capacity and enhancing partnerships among forestry information providers and users. (http://www.gfis.net/gfis/home.faces)
- International Union for Plant Information (IOPI) – a commission of the International Union of Biological Sciences which manages a series of cooperative international projects that aim to create databases of plant taxonomic information.
- Catalogue of Life (CoL) – which is a standardized dictionary containing the names of all known species. The NBII supported Integrated Taxonomic Information System (ITIS) partnered with Species 2000 in the creation of the Catalog of Life. (www.catalogueoflife.org/annual-checklist/2009)
- Census of Marine Life (COML) - the NBII is the U.S. node (OBIS-USA) of the Ocean Biogeographic Information System (OBIS). (www.nbii.gov/portal/community/Communities/Habitats/Marine/Marine_Data_(OBIS-USA)).

III. CORE COMPONENTS OF THE BIOLOGICAL INFORMATICS PROGRAM

Governments and institutions are calling for the collection of and access to a wide variety of critical information about the Nation's biological resources, including species, habitats, ecosystems, and potential impacts of natural and human activities [1]. The following core integrated components of the Biological Informatics Program have been created to design,

develop, implement, and update the capabilities needed to accommodate this mission:

- Gap Analysis Program - searchable data layers representing collections of digital species distribution maps, land cover, protected areas, and predictable habitat affinity models. Gap Analysis is a scientific means of assessing to what extent native animal and plant species are being protected. It can be conducted at a state, local, regional, or national level. The goal of Gap Analysis is to keep common species common by identifying those species and plant communities that are not adequately represented on existing conservation lands. Common species are those not threatened with extinction. By identifying their habitats, Gap Analysis gives land managers, planners, scientists, and policy makers the information they need to make better-informed decisions when identifying priority areas for conservation. Gap Analysis came out of the realization that a species-by-species approach to conservation is not effective because it does not address the continual loss and fragmentation of natural landscapes. Only by protecting regions already rich in habitat can one adequately protect the animal species that inhabit them.
- USGS Vegetation Characterization Program - The USGS-National Park Service (NPS) Vegetation Characterization Program is a cooperative effort by the U.S. Geological Survey (USGS) and the National Park Service (NPS) to classify, describe, and map vegetation communities in now hundreds national park units across the United States. The vegetation mapping program is an important part of the NPS Inventory and Monitoring Program, a long-term effort to develop baseline data for all national park units that have a natural resource component. This landmark program is both the first to provide national-scale descriptions of vegetation for a Federal agency and the first to create national vegetation standards for its data products. (www.nbii.gov/images/uploaded/8496_1181233610843_veg-factsheet.pdf)
- Integrated Taxonomic Information System (ITIS) - The first authoritative, standardized taxonomic reference information for the scientific and common species names of plants, animals, fungi, and microbes of North America (the U.S.'s only comprehensive, Web-accessible source of biological names) (http://itis.gov). The ITIS and Species 2000 Catalogue of Life (CoL) partnership provides the taxonomic backbone to the Encyclopedia of Life (EOL; http://www.eol.org).
- National Biological Information Infrastructure (NBII); Implements a full digital, interactive, distributed system that provides access to scientifically reliable natural science data and information of the Nation. (http://nbii.gov)

The data and information from the above programs and partner initiatives are passed through standards and protocols like ITIS and FGDC profiles to assure interoperability and quality. The NBII functionalities serve the needs of the user community through data and information access, integration and dissemination.

IV. NBII ENTERPRISE FRAMEWORK

The NBII Enterprise Framework, identifies the various Knowledge Management, Technology Management, and Program Management components of the NBII Program. The following list represents the primary Knowledge and Technology Management components of the NBII:

- NBII Web Catalog Resource Tool - The ability for NBII Nodes and partners to identify, catalog, and share resources quickly and seamlessly throughout the NBII network is supported through the implementation of this web-based Dublin Core metadata tool. (http://inputtool.nbii.gov)
- Web and Metadata Standards - The importance of a transparent user interface, regardless of where data exist or its format is vital in meeting your organization's data delivery and integration goals. The NBII currently has deployed a number of web standards, tools to implement and validate these standards, and provides technical support to NBII partners implementing these standards. (http://metadata.nbii.gov)
- OpenGIS Standards - The NBII fully supports the Open Geographic Information System (GIS) Standards as it relates to the sharing of web mapping applications and data layers. The NBII Open Viewer provides the capabilities for users to easily access distributed data layers from the over 35 web mapping applications currently deployed within the NBII network. (http://www.nbii.gov)
- BioBot Search Tool - The NBII BioBot Search tool easily facilitates access to all NBII data holdings, regardless of format or where they exist. The BioBot Search Tool serves as the overall integrator for NBII data and information within the network. (http://search.nbii.gov)
- My NBII Portal - The My NBII Portal is the primary means of collaboration, security, and system integrator within the NBII network. The My NBII portal services different communities of practice throughout the United States and provides a single user interface to support activities related to the NBII Program. (http://my.nbii.gov)

V. DATA STANDARDS AND TOOLS

The term bioinformatics includes the collecting and linking, storage, organizations, integration, analysis and synthesis, delivery, and application of biological data and information [1]. Through the development and/or application of bioinformatics,

the objective of NBII integration efforts is to bring together scientifically credible data and information in a manner that not only maintains attribution, provider ownership, and the integrity of data, but also adds value by applying standards and by providing further, similar data sources [3]. The NBII also develops new protocols, data processing, and other tools to assist in data integration. The NBII has multiple knowledge integration activities. The key activities are the cataloging of information sources, data harvesting from multiple servers and indexing them in one location, the hosting of collections, the coordination of experts, and the provision of platforms through information can be accessed. Specific examples of these tools include:

- Natural Resource Monitoring Partnership (NRMP) – provides Web-based tools developed by the NBII for the NRMP. These tools are for sharing natural resource monitoring project locations and a library of monitoring protocols. (http://nrmp.nbii.gov)
- NBII LIFE – provides a growing searchable collection of images related to nature and the environment. (http://life/nbii.gov)
- NBII Metadata Clearinghouse – provides standardized metadata-based descriptions of biological data sets and information products. It is a searchable collection of standardized metadata descriptions of biological data sets and information products based on the NBII's Biological Data Profile and the Federal Geographic Data Committee's (FGDC's) Content Standards for Digital Geospatial Metadata. (http://mercury.ornl.gov/nbii)

Metadata is a description of the content, quality, lineage, contact, condition, and other characteristics of data. The description of the data is organized in a standardized format using a common set of terms. Metadata is literally data about data. Metadata is a valuable tool as it:

- Preserves the usefulness of data over time by detailing methods for data collection and creation,
- Minimizes duplication of effort in the collection of expensive digital data and fosters sharing of digital data resources,
- Supports local data asset management, such as local inventory and data catalogs, and external user communities such as Clearinghouses and websites,
- Provides adequate guidance for end-use application of data such as detailed lineage and context,
- Makes it possible for data users to search, retrieve, and evaluate data set information from the NBII's vast network of biological databases by providing standardized descriptions of geospatial and biological data, and
- Makes information about data sets more easily accessible to scientists and researchers.

The NBII has found it very helpful to the user and/or data generator community to offer an extensive array of metadata resources. The NBII's Metadata Clearinghouse helps individuals easily search for and locate biological data and information from a variety of sources. The Clearinghouse is the biological node of the National Spatial Data Infrastructure (NSDI) Clearinghouse. The NBII also provides information and services regarding:

- Metadata standards (the NBII biological metadata standard has been developed as a "biological data profile" of the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata [CSDGM]),
- Metadata creation tools, and
- Metadata training workshops and resources for trainers.

The NBII Program has also developed a Geospatial Interoperability Framework (NBII-GIF or GIF) strategy based on International Standards Organization (ISO) standards and Open Geospatial Consortium (OGC) specifications, integrated with the core NBII Framework standards and protocols such as Universal Description, Discovery and Integration (UDDI) protocol, with data content and metadata standards.

The NBII Program represents a distributed set of Internet-based websites, resources, systems, services, databases, and applications. The purpose of the GIF is to allow users to seamlessly and transparently discover and visualize these distributed resources while providing the components, services and tools to its nodes and partners to participate in the framework. The goals of the overall GIF framework include the:

- Geospatial discovery of NBII's distributed content,
- Geospatial attribution of NBII content and metadata (not just the maps and data),
- Common infrastructure support for geospatial application development,
- Common infrastructure support for node management,
- Creation of sharable maps for collaboration across scientific communities, and
- Geospatial interoperability across mapping applications.

Lastly, the NBII has also developed the NBII Biocomplexity Thesaurus (BCT). The Biocomplexity Thesaurus displays terminologies and term relationships in the fields of biology, ecology, environmental sciences, and sustainability. Terminologies and term relationships in the fields of biology, ecology, environmental sciences, and sustainability. Now available via a Web service (SOAP). Development of the NBII Biocomplexity Thesaurus began in 2002-2003 through a partnership between the NBII and Cambridge Scientific Abstracts (CSA), a leading bibliographic database provider. The original Biocomplexity Thesaurus made available online in 2003, was a merger of five individual CSA thesauri: Aquatic Sciences and Fisheries, Life Sciences, Pollution, Sociological, and CERES/NBII. In 2004, the CSA Ecotourism Thesaurus was also merged into the Biocomplexity Thesaurus. In 2008, the Thesaurus was expanded to include new terms to support the fire ecology and management communities.

VI. NBII SUPPORTED SERVICES AND APPLICATIONS

The following list represents examples of actual NBII

supported application oriented capabilities for specific, often operational, scientific and natural resources management communities:

- Early Detection, Rapid Assessment, and Rapid Response (EDRR) to Invasive Species - A living catalog of resources for invasive species management. (http://www.edrr.nbii.gov)
- List of Invasive Alien Species (IAS) Online Information Systems - Internet-accessible databases and information systems providing species, bibliographic, taxonomic, expertise, distributions, images, and many other information types as they pertain to invasive, exotic, alien, introduced, non-native species, and all other species of the world's flora and fauna. (http://www.gisinetwork.org/Documents/draftiasdbs.htm#Toc212863847)
- Butterflies and Moths of North America (BAMONA) - A searchable database of verified records in the United States and Mexico including dynamic distribution maps, photographs, species accounts, and species checklists for each county in the United States and each state in Mexico. (http://www.butterfliesandmoths.org)
- Breeding Bird Atlas Explorer - A searchable database from Breeding Bird Atlases in North America used to assess the status of breeding populations of non-game birds at the state level. (http://www.pwrc.usgs.gov/bba/index.cfm)
- Fisheries and Aquatic Resources Data Access Wizard - A gateway to geospatial fisheries-related data; searchable using keywords, themes, geographic extent, and data providers. http://128.118.47.34/nbii/SearchPage.aspx?entry=FAR
- Monitoring Protocols Library and Monitoring Locator System - A perpetually evolving database developed by the Natural Resources Monitoring Partnership that provides access to monitoring protocols throughout the United States and Canada. (http://www.nbii.gov/portal/server.pt?CommunityID=819&spaceID=23&parentname=&control=SetCommunity&parentid=&in_hi_userid=200&PageID=0&space=CommunityPage).

VII. FUTURE NBII DEVELOPMENT

The NBII will continue to provide the premier enterprise system for accessing and disseminating integrated reliable biological information [1]. Overcoming the challenges will require a fully digital, Web-based, distributed system; scientifically reliable data and information; local to global data coverage; and effective decision making. The following goals have been established to develop the pathway forward:

- Increase the availability and usefulness of biological resources data and information (Content):
    - Obtain the broadest possible participation of stakeholders and partners in identifying high priority data and information needs.
    - Increase access to data and information for USGS Science Centers and partners.
    - Develop or select, implement, and promulgate standards and protocols to promote interoperability and information integration capabilities.
    - Increase the variety of data and information available through the NBII.
- Implement technologies and tools to integrate, analyze, visualize, and apply biological data and information to natural resource issues (Tools).
- Develop, apply, and promote the adoption of standard practices, protocols, and techniques to enhance knowledge discovery and retrieval form various resources (Infrastructure).
- Facilitate information science and technology research that supports the advancement of biological informatics capabilities (Research).
- Apply innovative technologies and best practices to improve development, description, and dissemination of biological information to customers (User Services).

REFERENCES

[1] M. Ruggiero, M. McNiff, A. Olson, and B. Wheeler, *Strategic Plan for the U.S. Geological Survey National Biological Informatics Program: 2005-2009*, U.S. Geological Survey, Biological Resources Discipline, Reston, VA, 2005.

[2] President's Committee of Advisors on Science and Technology, PCAST Panel on Biodiversity and Ecosystems, *Teaming with Life: Investing in Science to Understand and Use America's Living Capital*, Executive Office of the President of the United States, Washington, DC, 1998.

[3] P.B. Heidorn and A. Olson, "National Biological Information Infrastructure", *Encyclopedia of Library and Information Sciences*, 2010. Third Edition. DOI: 10.1081/E-ELIS3-120043271. In press.

[4] National Research Council, *A Biological Survey for the Nation*, National Academy Press, Washington, DC, 1993.

[5] R. Sepic and K. Kase, "The National Biological Information Infrastructure and an E-Government Tool", *Government Information Quarterly*, 19 2002, pp 407-424.