Pan-American Advanced Studies Institute (PASI)

Cyberinfrastructure for International Collaborative Biodiversity and Ecological Informatics

Context and Description of Proposed PASI (from the original funding proposal)

Biological diversity, or simply *biodiversity*, is the sum of life on Earth—plants, animals and microbes—encompassing all levels of biological organization from genomes to species to ecosystems. Approximately 1.8 million species are known as a result of 300 years of the biological exploration of the planet. Astonishingly, an estimated 15–50 million species await discovery and basic description. A grand challenge for the 21st century science is to expand and harness knowledge of Earth's biological diversity and to understand how it shapes the global environmental systems on which all of life depends. This knowledge is critical to science and society for rational policy for managing natural systems, sustaining human health, maintaining economic stability, and improving the quality of human life. The urgency for this knowledge increases daily as the conversion of natural systems to human-managed systems accelerates the decline of biological diversity at all levels of organization.

The importance of biodiversity research and education has been established by a series of landmark reports: U.S. NSF's *Task Force on Global Biodiversity* (Black et al., 1989), the systematics and biological collections community's *Systematics Agenda 2000* (1994), the Australian government's *The Darwin Declaration* (Environment Australia, 1998), Biodiversity II: Understanding and Protecting Our Biological Resources (Reaka-Kudla et al., 1997), and the U. S. President's Committee on Science and Technology's *Teaming with Life* (Lane, 1998).

Ecological and systematics research, which is aimed at elucidating the evolutionary history, geospatial pattern, community structure and ecosystem processes involved in the creation and maintenance of biological diversity, is becoming increasingly data-driven (Bisby, 2000; Bisby et al. 2002; Causey et al. 2004). Cyberinfrastructure is enabling this transformation, and as more primary sources of historic and real-time environmental data streams come on-line, the role of networked information services and collaboration technologies will continue to expand (Edwards, et al. 2000; Withey, et al. 2002). Biodiversity data is increasing at the rate 10^7 new records per year. Libraries, museums and research centers are grappling with the issues of dealing with the deluge of data in a stable, online and integrative way (Berman and Brady, 2005).

As environmental biology becomes increasingly data-driven, the nature of international collaborations in biodiversity will evolve to include more network-based interaction and communication (CIBIO, 2005). High-bandwidth networks continue to raise expectations for immediate access to research data sets and analysis tools, and more effective interaction with remote collaborators distributed around the globe. Expectations for essentially instantaneous access to data sets and to computational tools and to services will transform the timing and logistics of collaborative work. Requirements for scheduling field work in "seasons" will change

to requiring access to stored and streaming data in milliseconds. Data such as museum voucher records, remote sensing coverages, ecological experiment and real-time monitoring data, climate data, microclimate sensor data streams, DNA "barcode" sequences and phylogenetic trees, will all be in the mix of accessible and computable data types (Krishtalka et al. 2002; Herbert et al. 2003; Pennisi, 2005; Soberon and Peterson, 2004).

Simultaneously, networked sensor-based observatories in environmental biology are now clearly becoming the foundation for environmental monitoring of changes in biological diversity and other variables (Estrin et al 2001, 2003; Pottie and Keiser, 2000; Broad, 2005). The recent explosion of interest and work in planning the development of the US National Ecological Observation Network (NEON, http://www.neoninc.org) is a harbinger of the key role networked systems will play in streaming sensor data and other kinds of biological inventory information (e.g. observations and specimen voucher information) through web and grid services (Foster and Kesselman, 1999) for ecological research synthesis.

Projects such as the NSF-funded SEEK initiative (http://seek.ecoinformatics.org) are building architectures, protocols, and applications to integrate heterogeneous ecological data sets and to provide universal access to web browser based functions for scientific work-flow design for ad-hoc analysis.

Significance of Biodiversity Cyberinfrastructure Collaboration and Training

Cyberinfrastructure, such as research community computation and collaboration architectures, grid-based services, web services, automated semantic mediation for data integration, standard metadata profiles, standard information retrieval protocols and web-based research analysis applications, will transform the nature of much biodiversity and ecological research in the 21st century (Foster, 2005). The global nature of environmental issues and the international distribution of species and researchers demands regional, continental and ultimately global collaboration for biodiversity informatics training. Cyberinfrastructure will mediate much biodiversity research and monitoring collaborations, if the professional capacity is developed among researchers in many countries.

Central America is an excellent regional target for joint U.S. training in biodiversity cyberinfrastructure. There is strong historical legacy of research collaboration between Central American and U.S. universities and biodiversity research centers. The University of Costa Rica longest formal relationship with any U.S. university for international training of students is with the University of Kansas. Finally, the PIs recently organized an interdisciplinary workshop on the topic of CI for biodiversity research collaborations in Panama for this region, and we plan to leverage those contacts and interactions into this Study Institute.

Pan-American Advanced Studies Institute in Cyberinfrastructure for International Collaborative Biodiversity and Ecological Informatics in Costa Rica

We propose a Pan-American Advanced Studies Institute in *Cyberinfrastructure for International Collaborative Biodiversity and Ecological Informatics*. The aim of this PASI is to (1) expose Biodiversity Research and Biodiversity Informatics students from the U.S. and Central America to advanced concepts in distributed network-based science enabled by cyberinfrastructure tools, and (2) to promote a new organizational form for doing science, that is collaborative and

interdisciplinary. The training will be designed to enhance students with a strong biodiversity or ecology background with an empowering understanding of distributed computing and research network tools for collaborative research.

The variety of topics addressed and techniques used would be difficult to offer in a single university course. These topics will:

- emphasize the application of networked, cyberinfrastructure tools,
- focus on collaborative activities in biodiversity and ecological informatics in the U.S., Central America, Mexico and Colombia,
- include a curriculum involving network cyberinfrastructure, biodiversity and ecological informatics to give students an understanding of the role of each area and how they are interconnected, and
- demonstrate to students how they can advance a new organizational form for doing science, that is highly-collaborative, international and interdisciplinary.

Course Description

The course to be offered form a curriculum that will augment the student's knowledge and practical application of current cyberinfrastructure innovations in the areas that contribute to research and teaching in biodiversity and ecological informatics. The course will give students new data integration and analytical techniques with which to explore their discipline.

Invited lecturers will develop their short courses in more detail prior to the PASI; a listing and brief description of potential course subjects is given below. This information will be elaborated upon and posted on the PASI website. Lectures will be in English.

Potential Course Titles and Brief Descriptions

- **Grid Fundamentals and Application Frameworks:**
  Fundamental Grid Components and Technologies, Open Science Grid, Application Framework Concepts, hands-on examples from various grid-integrated science domains.

- **Web Services, Grid Services, Resource Frameworks:**
  Web Services and Grid Services concepts, Resources concept, Web Services Resources Framework concepts, Hands-On: Biodiversity web service examples, such as georeferencing services and web service integration with database software for cataloging collections.

- **Scientific Workflows, Data Integration, Semantic Mediation:**
  Biodiversity eScience concepts, scientific workflows, process management, coordination, Hands-On: Workflow processing from data acquisition to research results using Kepler or Taverna.

- **Biodiversity Data Conceptualization and Management:**

Representation of information from natural systems, data heterogeneity, scaling, encoding, transformations, specimen and observation system, ontologies, databases, versioning.

- **Metadata-based Ecological Dataset Description, Discovery, Retrieval and Integration:**
  Biodiversity tools for dataset description, indexing and cataloging, discovery, search and retrieval protocols and tools for structured biodiversity and ecological data. Utilizing applications such as **Metacat** and **Morpho,** and tools based on TDWG standards for description for collections and observations.

- **Biological Collections Data Description, Discovery, Integration and Retrieval:**
  Software tools for specimen and taxon data description into databases, **DiGIR**, **BioCASE** and **TAPIR** protocols for specimen and observation data retrieval and integration.

- **Biodiversity Data Visualization:**
  Classification and collections data visual analysis tools, network-based retrieval and caching, web browser and applet options. Lecture and laboratory based on various existing software applications.

- **Environmental Sensor Data Management:**
  Lectures and Field site Hands-On: hardware and software tools and interfaces for capturing sensor data from networked sensor platforms based on the UCLA Center for Embedded Networked Sensing, Networked InfoMechanical Systems.

- **Biodiversity Informatics Networking:**
  Lectures on: current networking initiatives that aim at facilitating the digitization, management and delivery of biodiversity information (e.g., GBIF, IABIN), critical technical, social and biological issues in networking. Hands-On: experience using biodiversity information systems about Costa Rican biodiversity (*Atta* system developed by INBio) and from other countries (GBIF portal).

Laboratories will be structured to show students realistic examples of biodiversity research and biodiversity informatics using the cyber tools. Laboratories will be paired with lectures, linking concepts and constructs with methods and artifacts, giving students the opportunity to understand the relationships between the practice of the science and the application of cyber tools. Lab exercises will be guided in English and in Spanish.

U.S. Organizers

James Beach, University of Kansas, and Julio Ibarra, Florida International University, are the PIs for the PASI, and co-chairs of the Organizing Committee. Beach, Director of Informatics at the Kansas University Biodiversity Research Center and an active biodiversity informatics researcher, is co-lead for organizing the biodiversity research and bioinformatics component of the PASI. Ibarra, Executive Director of the Center for Internet Augmented Research and

Assessment (CIARA) at FIU and recipient of the NSF International Research Network Connections (IRNC) grant award for Latin America, is co-lead for organizing the cyberinfrastructure component of the PASI.

Beach and Ibarra co-organized a January 2006 workshop on Cyberinfrastructure for International Biodiversity Research Collaboration[1], in Panama (NSF/OISE award #0549456) that brought together domain scientists, practitioners, policy makers and funding agency representatives to discuss the issues and challenges of building and sustaining technology infrastructure needed for international biodiversity research collaborations. The workshop focused on identifying activities that will enable North and Central American partnerships, with an eye toward a more inclusive western hemisphere approach. Outcomes will include a report with recommendations for follow-on activities and suggestions for investments to promote international informatics collaboration.

Costa Rican Organizers and Hosts

Costa Rica is the preferred host country, because of its leadership role with biodiversity research, inventory and informatics in Central America, and because of its higher education relationships with U.S. institutions. The local organizing organizations will be the Universidad de Costa Rica (UCR), INBio and the Organization for Tropical Studies (OTS).

For almost 50 years, scientists from U.S. universities have forged collaborative working relationships with colleagues at the University of Costa Rica in the interest of strengthening education and research in tropical biology. The University of Kansas-Universidad de Costa Rica student exchange program, initiated in 1958, is the oldest inter-university exchange of its kind in the western hemisphere. It includes education in the biological sciences.

INBio, the National Biodiversity Institute of Costa Rica (www.inbio.ac.cr), is a research institution that has been internationally recognized as a pioneer in Biodiversity Informatics. It conducts an exhaustive inventory of Costa Rican biodiversity that has led to the development of sophisticated software tools to capture, manage and disseminate digital information about Costa Rican biodiversity. INBio is one of the leader organizations in international biodiversity initiatives such as GBIF (Global Biodiversity Information Facility), IABIN (Inter American Biodiversity Information Network), SIAM (Sistema de Información Ambiental Mesoamericano) and EoL (Encyclopedia of Life). Jointly with the ITCR (Instituto Tecnológico de Costa Rica) it recently established the first UNESCO Chair in Biodiversity Informatics. This chair, which has a Latin American scope, aims at developing graduate level training courses in this area and at consolidating more than five years of experience in biodiversity informatics training at INBio.

OTS, with its primary offices at Duke University and on the UCR campus, was founded in 1963 by a consortium of universities to provide leadership in biological education and research in the tropics. OTS owns three biological stations in Costa Rica and offers several highly-regarded

[1] http://www.ciara.fiu.edu/biocyber/index.htm

graduate and undergraduate-level courses each year in tropical biology. OTS is ideally situated and experienced to handle the travel logistics for the Institute.

Venue

The main venue will be the La Selva Biological Station[2] of the Organization of Tropical Studies (OTS).

Organizing Committee

The Organizing Committee will consist of researchers and practitioners in biodiversity and cyberinfrastructure from the U.S., Costa Rica, Mexico and Panama. The Organizing Committee will develop the curriculum and select the students for the Institute.

Organizing Committee Members

- **Peter Arzberger**, Chair of the Pacific Rim Application and Grid Middleware Assembly (PRAGMA), an open, institution based organization, consisting of 25 institutions around the Pacific Rim. Arzberger will provide coordination with the international cyberinfrastructure community.

- **Gisele Didier**, Coordinator at the Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT), Panama. Didier, will provide coordination with the biodiversity research and cyberinfrastructure community of Panama.

- **Bryan Heidorn,** University of Illinois, Graduate School of Library and Information Science. Heidorn has broad interests in informatics, including **t**he expansion of the functionality of automatic methods for semantic markup, natural language understanding, spatial language, information extraction; human-computer information system design, ecological information management.

- **Phil Rundel, William Kaiser, and Eric Graham**, UCLA Center for Embedded Networked Sensing (CENS). Kaiser, Rundel and Graham will organize the sensor data network training activities. They are currently working at La Selva to install sensor arrays and a Wi-FI network on a recently funded project sensor network project.

- **Atzimba Lopez**, Mexico Long Term Ecological Research Network, Mexico. She will provide coordination with the biodiversity and ecological informatics community of Mexico.

- **Erick Mata**, Associate Director of INBio, Santo Domingo de Heredia, and Associate Professor at the Institute of Technology of Costa Rica, Cartago. Mata's research areas include: genetic algorithms, biodiversity informatics, data visualization, object oriented systems, multimedia systems and algorithmic graph theory. He will provide coordination with the bioinformatics and cyberinfrastructure community in Costa Rica.

[2] http://www.ots.duke.edu/en/about/

- **Robert Morris,** Computer Science, University of Massachusetts, Boston. Morris is an active researcher in biodiversity informatics working in areas of XML mediation and integration of species data sources, biodiversity ontologies, and with the development of species description standards with TDWG and GBIF, among many other areas.

- **Deanna Pennington**, University of New Mexico, Albuquerque, NM. Pennington is an active ecological informatics researcher and educator. She is involved in several eco-informatics research efforts including the US NSF-funded SEEK Project, and an NSF CI-TEAM project to develop effective collaboration processes for engaging scientists in biology, computer science and informatics in effective interdisciplinary collaborative projects. She has organized tutorials and workshops on new technologies for semantic mediation of disparate data sets, for work flow tools, and various other related topics of ecological data integration and analysis.

- **Mark Schildauer**, National Center for Ecological Analysis and Synthesis (NCEAS), UC Santa Barbara. Schildhauer is developing and supporting advanced technological solutions for ecologists to collect, store, access, and analyze data. His work explores how scientific workflows and formal ontologies can create powerful GRID-based analytical frameworks that facilitate replicating scientific results, and sharing of both data and code.

- **Ana Sittenfeld**, Director of the Office of International Affairs and External Cooperation (OAICE), Professor of Microbiology at the Center for Research in Cellular and Molecular Biology (CIBCM), at the University of Costa Rica. Dra. Sittenfeld and her associates, Dra. Gabriela Marin and Dr. Vladimir Lara of the Escuela de Ciencias de la Computacion e Informatica, UCR are also contributing to the PASI planning.

- **Jorge Soberon**, Biodiversity Research Center, University of Kansas, Lawrence. Soberon's areas of expertise are theoretical population ecology, conservation biology and informatics for biodiversity. He has extensive experience in the biodiversity of Mexico, and is currently exploring novel technical approaches for the use visually exploring the data parameter, geographic and phylogenetic spaces associated with ecological niche models.

Criteria for Lecturers and Students

Lecturer Selection Process

Lecturers for the Institute program shall be identified with the help of the Organizing Committee members and from the contacts established from the recent Workshop on Cyberinfrastructure for International Biodiversity Research Collaboration, in Panama. The goal will be to select researchers and practitioners who are active in the fields of biodiversity research, bioinformatics and cyberinfrastructure who are committed to participate in the Institute and who are interested in fostering an international collaborative biodiversity research program.

Student Selection Process

Students selected will be recognized by their home institutions and colleagues to be of the highest caliber. An equal number of students from Latin America and the United States will be selected. The budget and justification is based on participation of 25-30 students in total.

The student selection process will begin by announcing the institute through direct mail, project- and domain-specific lists for biodiversity research, bioinformatics and cyberinfrastructure. A call for proposal announcement mechanism will be used whereby students and faculty will submit a one-page proposal explaining how participation in the institute will benefit their research and pedagogy. The students will be asked to describe what the scope of impact will be for them, if they participate in the course. The applications will be judged by the organizing committee.

Web-based Publication and Dissemination

A web site will be created to provide up-to-date information on the PASI, with specific details on the activity, including recruitment procedures, meeting topics, links to related activities. The web site will be used for disseminating information before and after the PASI.

Curriculum information on the courses will be posted on the Institute web site. The web site will provide registration form, a participant list and logistics information about hotel and travel information to assist both student participants and instructors. Furthermore, the web site and mailing list will be a critical support mechanism for the organizing committee to select appropriate students and post their participation proposals. Participants will use the web site to keep in touch before and after the institute. It will be especially useful for students in Latin America to contact the instructors or other students if they should need to, and for all participants to have easy access to the written materials.

Schedule of Courses and Laboratories

This Institute program will be for a a period of 10 working days. The Co-Organizers will be present for the entire Institute program, but the instructors may attend for a shorter period. Lecturers may be present for shorter periods.

The curriculum will be a mix of lectures, seminars and labs. Lectures will be typical faculty-level presentations of an informatics research or infrastructure deployment topic with an emphasis on science objectives and biology research collaboration. Seminars will be structured more as interactive tutorials with a presentation linked to short hands-on exercises and demonstrations. Lab periods will be completely hands-on with structured assignments with specific tasks to be accomplished by the students. The curriculum is structured in a uniform manner, allowing time for informal discussion and professional socializing in the late afternoon and evening.

Intellectual Merit

The PASI will address biodiversity and ecological informatics, a research discipline that is building the cyberinfrastructure foundations for environmental research in the 21st century. CI-based research collaboration will likely be the predominant mechanism for investigations in the environmental sciences in the future, as is now occurring in the physical science communities,

such as astronomy and physics. The NSF BIO directorate recently published a report on the importance of cyberinfrastructure for research collaborations in biology (CIBIO, 2005.)

Broader Impacts

The PASI we propose is focused precisely on the broader impacts of engaging biodiversity informatics researchers in the U.S. and Central America to deliver an intensive training exercise in technologies related to research collaboration for Master and Ph.D. students. The PASI is seen as the next step in a set of international activities to establish new research and training collaborations among the participating nations. The ultimate broader impact of this activity is to enable biodiversity researchers to effectively collaborate on addressing the most vital environmental science issues of our time.

###