Research Agenda for Biodiversity and Ecosystem Informatics (BDEI)

Judy Cushing (The Evergreen State College)
Kate Beard-Tisdale (University of Maine); Kathleen Bergen (University of Michigan)
Jim Clark (Duke University); Geoff Henebry (University of Nebraska-Lincoln)
Eric Landis (Natural Resources Information Management)
David Maier (OGI School of Science and Engineering); John Schnase (NASA)
Rob Stevenson (University of Massachusetts, Boston)

The United States faces biological and ecological threats, e.g., West Nile virus, invasive species, drought, biological warfare agents, and global change (Lubchenco 1998, Wilson 2002, Jenkins 2003, Karl and Trenberth 2003, and State of the Planet Series 2003). At the same time, new sensors – aloft, among, and in situ – yield more data than ever before, and the internet brings unprecedented access to both fresh and archived data. These problems and opportunities require multi-threaded, complex solutions – deeper theoretical understanding of the underlying ecological systems, better retrospective and current data, better resource management and policy decisions, and improved response capabilities.

While the computer science and eco-informatics research communities have begun to contribute to solving these problems with applications such as the USDA plants database (http://plants.usda.gov), *Lifemapper* (http://www.lifemapper.org), *Morpho* – a metadata editor using the Ecological Markup Language (http://www.ecoinformatics.org/tools.html), and the Science Environmental for Ecological Knowledge (SEEK) project (http://seek.ecoinformatics.org), much remains to be done.

To determine the research requirements of a new ecosystem informatics, the National Science Foundation (NSF), U. S. Geologic Survey (USGS) and National Aeronautical and Space Administration (NASA) have collaborated to elucidate what eco-system informatics tools are lacking and what research is needed in order to build such tools. Results presented in this poster are based on presentations made at a February 2003 workshop[1] by principal investigators of an NSF pilot program in Biodiversity and Ecosystem Informatics (BDEI) reported to NSF-NASA-USGS research directors. This poster will summarize research results from 12 NSF-funded ecosystem informatics projects, and outline future work as identified by those principle investigators. We have organized the BDEI projects into four research categories as follows:

1. **Semantic Data Integration.** Data needed to address critical questions in ecology are scattered, heterogeneous, and complex. Large volumes of diverse data types must be semantically integrated. Early integration efforts generally assumed that terms and formats (syntax) and meaning of terms (semantics) were the same. In the web environment and in the decentralized, heterogeneous information collections relevant to ecological research such simplifying assumptions do not hold. Research approaches for information integration include thesauri, ontologies, and metadata services. Existing digital gazetteers (USGS GNIS and NIMA GNS) support spatial integration, but need further development. Agent architectures show promise for linking ontologies with various metadata services.

2. **Spatio-temporal Data & Sensing Technologies**. A recent confluence of technologies – orbital sensors complemented by mobile teams of citizen-scientists and specialists equipped with geo-referential proximal sensors and robust wireless networks of reactive sensing agents – result in data streams rich in dimensionality and across a wide range of spatial, temporal, spectral, radiometric, thematic, and taxonomic scales. Robust representations capturing the complexities of environmental patterns and processes are required. Flexible database schemata and sophisticated queries, algorithms for spatio-temporal analysis, wireless reactive agent networks and tools for metadata acquisition,

[1] See http://canopy.evergreen.edu/bdeipi.

management and interpretation are the most critical research areas. Enhanced sensory presentation of environmental data encompassing visual, sonic, and haptic feedback systems are needed to explore high-dimensional data spaces.

3. Modeling and Forecasting. Advanced frameworks, including hardware for adaptive and intelligent systems and high performance computing needs are needed. Memory and speed continue as important bottlenecks, so coupling of models, modular models, and hardware grid and distributed computing are needed. Model-Data interactions key since many ecological problems could be informed by existing large data sets, but methods are needed for inference of high dimensional problems. Data are indirect, massively unbalanced with missing observations, and subject to many sources of stochasticity. Processes interact at a range of scales and hidden 'parameters' are often more like variables. Data exhibit variability, uncertainty, and complexity. More than just "pictures", meaningful visualization must be part of working models on which measurements may be made and change simulated. Biodiversity science requires more efficient algorithms, ways to deal with incomplete knowledge, better understanding of spanning spatio-temporal scales, and techniques for high dimensional problems. Derived data products and provisioning data at multiple places are also critical.

4. **Putting it into Practice – the human-side of ecosystem informatics**. Advancing an eco-informatics agenda hinges on closing various "digital divides", shaping an eco-informatics culture and developing socio-technological partnerships. We must solve social and ethical conundrums among various players: industry and science; bioinformatics and eco-informatics; ecologists conducting "big environmental science" and traditional individual ecologists working in relative isolation; scientists and information managers; professional and citizen scientists; and the tripartite scientists, resource managers and policy makers. R&D budgets of telecommunications and consumer electronics companies as well as the defense and the health services industries, and even bioinformatics, dwarf eco-informatics budgets; BDEI could look to those areas for technologies to adapt and transfer. Stakeholders should adapt open source models to enabling technologies. Finally, just creating the technology is not enough. Even with the needed technology, extensive user training will be required. In addition, questions such as who should contribute information, who should have access and how should information be shared, and how many research resources should be allocated for information management, all need serious consideration.

Four research areas cut across the above categories: 1) Semantics, metadata, and data provenance support, including terminology management with ontologies and glossaries, attaching metadata to data, and other data annotation; 2) Adaptive, flexible database schemas, domain-specific data types and schema management, including specialized support for spatio-temporal data and support for different data models; 3) Data acquisition, documentation and cleaning, including retroactive data capture to speed up digitization of data from static media, *in situ* and remote sensing technologies, all with automated or semi-automated metadata acquisition, updating, and management; and 4) Modeling and analysis infrastructure, including better mathematical and statistical models of organisms and ecological and biological systems, better visualization, integration of data models with models, uncertainty quantification, and missing data management.

BDEI principal investigators further emphasize that social and cultural contexts of technology transfer are critical, and that future eco-informatics research must be considered in terms of a variety of researchers, students, resource and information managers, and policy makers (including many agencies, such as the US Forest Service, US Fish and Wildlife Service, Bureau of Land Management, National Park Service, and the Environmental Protection Agency, as well as state and local resource agencies), and. In addition, funded eco-informatics research artifacts must be couched in the context of today's web-enabled research and learning communities, i.e., provide the cyber infrastructure needed for tomorrow's ecologists.