The Future Of Ecoinformatics In Long Term Ecological Research

James W. Brunt

Department of Biology, University of New Mexico

Albuquerque, NM 87131-1091

and

Peter McCartney

Center for Environmental Studies, Arizona State University

Tempe, AZ 85287-3111

and

Karen Baker

Scripps Institution of Oceanography, University of California San Diego,

La Jolla, CA 92093-0218

and

Susan G. Stafford

Department of Forest Sciences, Colorado State University

Ft. Collins, CO 80525

ABSTRACT

Emerging information technologies allow new exploration into tools for the management and use of information that solve problems for ecologists and create new and innovative lines of scientific inquiry. Collaborative, multi-disciplinary research programs to facilitate these new lines of inquiry have produced a need for scientific information systems that communicate data, information, and knowledge across spatial, disciplinary, and cultural boundaries.

INTRODUCTION

Increased need for ecologists to examine global change, bio-complexity, and sustainability is resulting in research and synthesis at larger spatial and temporal scales than traditionally addressed in ecological studies. The development of collaborative, multi-disciplinary research programs has produced a concomitant need for scientific information systems that communicate data, information, and knowledge across spatial, disciplinary, and cultural boundaries. The primary motivation for developing scientific information systems must be the new types of scientific inquiry that they make possible. Information systems science and related information infrastructure are leading to a paradigm shift in biology [1][2]. Thus far this has been most evident in the genomic community [2], where the creation of databases and associated tools have facilitated a tremendous increase in the understanding of the relationship between the genetic sequences and the actions of specific genes. Ecology is perched on the brink of a similar expansion, brought on through improvements in software tools and data communication. In long-term studies, retention and documentation of the data are the foundation upon which the success of the overall project succeeds or fails. Long-term studies also depend on information systems to facilitate sharing of data and to combine data for the purpose of integrated multidisciplinary projects. In addition, public decisions involving environmental policy and management frequently require data that are regional or national, but most ecological data is collected at smaller scales. Information systems make it possible to integrate diverse data resources in ways that support decision-making processes.

We recognize that knowledge (in the broad sense) is generated through an iterative process of acquiring data, transforming it into useful information, and drawing inferences that enable us to achieve understanding and informed make decisions. Information management is slowly undergoing an evolution through these three domains.

The Long Term Ecological Research (LTER) Network is a collaborative effort involving more than 1100 scientists and students investigating ecological processes over long temporal and broad spatial scales.

- The Network promotes synthesis and comparative research across sites and ecosystems and among other related national and international research programs.
- The National Science Foundation established the LTER program in 1980 to support research on long-term ecological phenomena in the United States.
- The 24 LTER Sites represent diverse ecosystems and research emphases.
- The Network Office coordinates communication, network publications, and research-planning activities.

As LTER moves into its third decade, ecoinformatics will continue to play a critical role in defining and facilitating this

expanding, new ecology. The third decade of the LTER has been designated as a *Decade of Synthesis* for which the scale and complexity of the information management tasks presents a number of challenges for organizing and coordinating the diverse skills and resources within the LTER Network of research sites.

The LTER program was designed from its inception to incorporate data management as an integral component of the research. Within the overall goal of facilitating research, data managers at the site-level spent significant portions of their effort in developing methods to handle documentation and the custodial aspects of codes, data formats, and consistency during the 80's. In the 90's we built upon these developments, utilizing the rapidly expanding technology of the Internet to address the design of information systems. As we move to the next millennium, our goals are now expanding to building an infrastructure for the next level, a parlaying of information management into knowledge management.

In addition to supporting LTER site and intersite research, the data management program within LTER provides information to a diverse set of end users who expect to have access to our publicly funded research data. As scientific inquiry becomes more multi-disciplinary, we are challenged to find solutions for making primary ecological information more directly useable by (and valuable to) non-specialists. We embrace the research goals as a guiding force behind our work, but we also recognize that general scientists, public policy makers, businesses, the legal profession, K-12 educators, and even the entertainment industry, all make use of information about the natural world. There can be serious implications if those users make uninformed decisions based on faulty, outdated, or incomplete information.

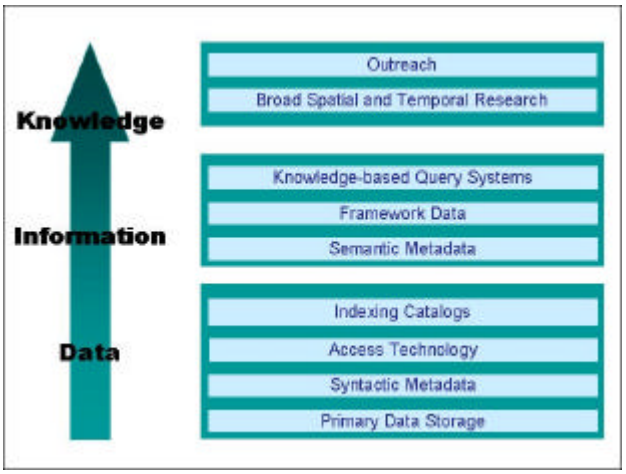
The purpose of this paper is to summarize the directions that information management, both within and outside the LTER program, need to look to respond to the changing needs of ecological science as we enter the next millennium.

METHODS FOR ADDRESSING ECOINFORMATICS CHALLENGES IN THE ECOLOGICAL SCIENCES

The challenges and expectations identified above call for the application of information technology beyond simple data storage and electronic publication to development of an active, globally integrated information network with the capacity to discover, access, interpret and process data facilely across the comparability and scaling barriers. Creation of this infrastructure requires investment of effort and resources into three broad areas that span the transformation of observations from data to information and to knowledge (Figure 1):

- Design a system of networked data storage to provide long-term management and accessibility of ecological data.
- Develop tools and procedures for facilitating the integration and synthesis of primary data.
- Promote and support research activities focusing on applications of archived data sources in broad, synthetic research.

Figure 1. Component Model of Ecoinformatics Infrastructure



DATA: Establishing Data Storage Network

We foresee a series of activities contributing to development of a global infrastructure for ecological data management and access. We envision this infrastructure developing through (1) the creation of data repositories that actively accumulate valuable data sets and ensure their long-term viability and accessibility, (2) development and adoption of standards for documenting databases and the research that produced them to enhance the usability of these data, and (3) development and adoption of procedures that will reverse the traditional attrition trends for data sets.

Long term management:

Data management plans seek to ensure data quality and availability. Quality control protocols operate from field sampling, lab analysis to data management. In the data flow, data quality is maintained through application of a variety of quality assurance/control procedures such as foreign key enforcement, validation triggers, and exploratory data analysis (EDA) screening.

Effectiveness of these methods is determined by the extent to which data management plans can be included in the actual research design for data collection. Availability of data can be threatened by both short-term factors such as power or device failure or long-term factors such as technology drift, media decay or format obsolescence. Solutions for these difficulties involve regular backups to removable media, and/or use of redundant subsystems. Long-term availability is dependent on maintenance of data in an online system with a plan for timely migration to new hardware and software. To successfully maintain data, it is necessary that an institution have the resources for and the commitment to managing and upgrading the necessary equipment and software, and to maintain connectivity. While it is relatively easy and inexpensive to install a data server and a web connection, it is quite another matter to keep that connection up over several years, let alone in perpetuity. Both site resources and commitment are required to form such a network of data repositories.

The most familiar type of primary data in ecological sciences is tabular data. Storage formats such as SQL and object-oriented

databases continue to evolve following open standards set by international consortia. New technologies are enabling far more sophisticated solutions for text data such as "grey" literature reports or records. With Extensible Markup Language (XML), it is possible to go beyond simple indexing of electronic documents to actually mapping their structure, enabling more rapid and accurate location of relevant information. For example, instead of searching the index for every occurrence of the words *bird* and *populations* in the hope of finding some survey data, a search engine could locate the section of a document tagged as <census results> to access the information requested. With XML rapidly being adopted as the language of the internet during this remarkable time of transition, it holds the potential to fill a variety roles in both the management and the exchange of information in the future.

Access technology

Most LTER sites now possess technology to provide access to data through a variety of methods. These range from simple downloads of static files to interactive query applications that support more sophisticated search and selection of information. Selected LTER sites have been active in exploring new approaches to data management. The Z39.50 was adopted as a client server search technology and is in common use among libraries and museums, as a platform-independent capability for searching multiple data sources on diverse hardware and software implementations. Further, the LTER Network Office, in partnership with the San Diego Supercomputer Center are considering the Storage Request Broker (SRB), software providing a single interface to hide differences between data storage systems. Many sites are active with geographic information system (GIS) so are poised for activity on map-based query interfaces which can provide a spatial framework for referencing data queries.

While these approaches are subject to continued innovation, a recognized limitation of some of these solutions is that they are inherently proprietary to the specific data content, storage and delivery system and thus are time-consuming to develop. A layer of open access technology needs to be draped over this network of data repositories that could facilitate the ability to conduct the most fundamental search and query operations from a single agent using a single protocol. Several technologies currently in development and limited implementation can provide multi-tier solutions:

- XML (eXtensible Markup Language) - a universal language for data exchange that is self-documenting,
- UDDI – Universal Data Discovery and Integration
- WSDL Web Services Description Language

Syntactic (Data-bound) Metadata

The term metadata refers to data that describe data. Metadata represent the key elements to transforming archived data sets into useable research resources. In this level of the model (Figure 1), we are concerned with information about the syntax of the data – information that describes each specific data set. This information is inextricably bound to the data set and is thus expected to be stored and managed in close conjunction with the data.

In a seminal paper about the survival of ecological data, Michener et al. [3] identify five levels of metadata description required to fully document an ecological dataset. These range from information about the research project that produced the data – names of investigators, sampling strategies, collection methods, etc. – to detailed attributes of the columns, data types and file formats of the data tables that were archived. The Federal Geographic Data Committee has published a standard for the content of spatial metadata that exhibits a similar hierarchical, albeit more exacting content structure. A Kansas University effort has developed a metadata standard for indexing and describing museum collections data; similar efforts to index electronic resources on the World Wide Web have been made by the Dublin Core (http://purl.org/dc).

There is a need for the development of widely accepted standards for ecological metadata that go beyond this simple beginning. For this to become a manageable task, these metadata need to be developed following a modular approach similar to other well know standards efforts such as the W^3 Consortium, FGDC and Dublin Core. Discrete working groups focusing on specific content domains within ecological research would produce modules that would be required only for datasets that contain certain kinds of information. Each product of these workshops would contribute toward a comprehensive standard that serves not to dictate research methods, but rather how to effectively document the structure and design behind one's methods and observations.

Indexing Catalogs

As the corpus of online data resources grows, the need for efficient indexing and searching far outstrips the capacity of static, unsophisticated aids such as html link pages and webcrawler-based search engines. Current efforts to develop online metadata catalogs such as the LTER Data Table of Contents database (URL: http://www.lternet.edu/DTOC) are building a valuable infrastructure for navigating the growing network of digital data. The design of these catalogs should be sufficiently open to support searches by search applications commonly in use across the internet such as Z39.50 and follow a development model based on other indexing efforts such as the National Spatial Data Infrastructures network of clearinghouses for geospatial data sets, the Council for the Preservation of Anthropological Records (http://archaeology.asu.edu/copar), and the Dublin Core.

Sustainability

For a system to be sustainable, a strategy is necessary for handling the incorporation of new data into the knowledge network. Despite its vital importance, few active research projects have the time or take the effort to produce metadata for their research data is often prohibitive; another barrier is that until recently no adequate guidelines have been developed[4][5]. Practices observed in other disciplines suggest several options. One is to encourage funding and permitting agencies to endorse the submission of research data into knowledge repositories and to adopt a set of standards for this process. Another is to work with professional societies to develop programs that create recognition for data archiving and documentation. One such partnership, created through cooperation between the Ecological Society of America and the San Diego Super Computer Center, developed a peer-review

process for datasets and associated metadata, with successful submissions being published in the ESA journals. Finally, the cost and difficulty of creating metadata might be mitigated by development of freely distributable tools that automate the documentation process through reverse-engineering of data files and use of "wizard" forms that query the investigator for information similar to the way tax preparation software gathers financial backgrounds.

INFORMATION: Integration and synthesis of data

Information is modeled here as a bridge between primary data and knowledge. We recognize a series of technologies and activities that create the interface between data storage systems described above and the kinds of synthetic research questions we wish to accommodate within our broad infrastructure. One component is a set of standards for decomposing these questions into smaller elements that can be documented in standardized, machine-parsable form. Another is a very experimental approach involving the development of *expert systems* - sophisticated software tools that are capable of performing intelligent searches and processing of diverse datasets. The third, enabled in part by the second, is to identify a basic set of parameters relevant to the most broad inquires in biodiversity and produce synthetic framework data in the form of GIS covers and/or summary tables.

Semantic (Query-bound) Metadata

Comparable metadata necessary to perform translation and processing for scale matching needed to forge compatibility between two datasets based on their metadata descriptions. This latter body of metadata we refer to as semantic metadata because it concerns itself not with the organization of information but with its meaning. It is query-bound in that it documents our diverse units of inquiry, not just the more familiar syntactic metadata that documenting diverse units of observation.

This class of metadata might consist of calibration curves, thesauri that equate nominal categories, or machine-readable codes defining a particular set of processing steps required for certain types of data. This information might be included with the metadata of the particular datasets or be stored in a separate, central knowledge base to which regular additions are made. As a simplistic example, metadata associated with two different datasets might indicate that radiocarbon dates from one palynology core giving spore/pollen counts were reported in calendar years calibrated to a particular curve while those from another were reported as raw dates. Reference to a knowledge base containing the necessary calibration data and the appropriate processing steps would enable an expert system to retrieve the data and cast them in a compatible form by cross-calibration. As is the case for this example, the knowledge and software tools for performing many of these processing steps already exist. Development of an expert system in such cases will involve gathering and electronically encoding the various calibration curves, thesauri, classification rules, etc. used in existing databases and either developing or incorporating existing processing code that can act upon this information.

Develop tools and procedures for automated integration of data

Much of the existing technology for query and retrieval of data requires some degree of familiarity with the data structure and its meaning. However, few tools exist to facilitate the task of synthesizing data from diverse primary sources. We may expect that the not-too-distant future will bring sophisticated search engines and software tools for automating many data synthesis steps presently done by hand. These tools would be based on technologies such as expert systems and would be able to (1) respond to some relatively easy-to-use query language, (2) access both semantic metadata about the categories of information requested and syntactic metadata about the databases to be searched, and (3) perform a certain amount of query, evaluation, and processing of primary data prior to returning a result.

Within an overall strategic plan, we expect the initial products of this effort would be a set of loosely integrated software components that will accumulate and evolve as new information is brought into the system. It is untenable at this point to envision a single massive system with a single interface to respond to all queries against all known data sets. However, the tools exist today to develop some discrete components that can automate many of the routine and time consuming tasks associated with synthesizing diverse data sources.

Expert systems exist and are in common use in other fields. The key to developing these tools lies in two areas. The first involves extending our partnerships with expertise in sophisticated computer technology such as artificial intelligence, expert systems and neural networks. The expertise is not traditionally found within ecological sciences nor are the funding sources for these disciplines familiar ground for ecologists. The other area will involve more in-house effort. It will be necessary to adopt a language for encoding both syntactic and semantic metadata in machine-readable form. This form will likely be based on a language such as Resource Description Format (RDF), an XML based standard for encoding machine-readable metadata under development by the World Wide Web consortium. XML, like other SGML derived languages provides a rich syntax for expressing complex, structured information and has extensive support in the commercial industry.

Pilot projects such as the LTER Network Information System (NIS) [6][1] provide a means of mobilizing collaborations for combining research questions with technology and algorithms. These projects are typically small enough in scope not to require extraordinary resources for implementation, and just complex enough to demonstrate the scientific principles and the technological methods used. While simplified in detail, they approximate what a full-scale, full-complexity implementation would be like, even if the prototype turns out to be neither transportable nor scalable.

The demands this sort of approach places on the development of metadata are significant and justify continued investments in developing metadata standards and in streamlining the process by which metadata are generated. To be effective, existing efforts will need to be augmented with more rigorous procedures for encoding semantic metadata such as classification systems, measurement parameters, analytic procedures, etc. Current metadata implementations still rely

heavily on open text representations of information and thus will require further revamping to meet the rigors that this sort of advanced processing tools require.

KNOWLEDGE: Promote and support research collaborations integrating information at broad spatial or temporal scales

Funding strategies for information technology during the last two decades have focused on developing infrastructure with the notion that "if you build it, they will come". The new crop of initiatives shows much more concern with ensuring that our data products are of significant value to current and future research – that is, new proposals are expected to provide *application*, not just availability, of data. Like most current research efforts, construction of this broad information infrastructure must have a strong question-driven component to enable a process of feedback between end-users and data sources. We need to challenge the ecology community to provide recognition to archive and share their data so that we can create both incentives and guidelines for developing the infrastructure components outlined above.

To bring about these changes requires adopting new perspectives on the relationship between data management and research. The traditional approach is for ecological scientists to make use of selected technology (and computer science methods). Recently, however, we see evidence from the development of database technology, web technology, modeling and simulation techniques, that advances in computer science and technology can provide new methods that can influence how ecological science is conducted. These are new opportunities for scientific explorations that were not conceived by ecologists. Also these advances are not directly driven by the current practice of ecologists. That is, there are opportunities for computer science to suggest changes (advances) to the methods of ecologists. LTER partners recognize that challenging problems in ecology provide focus and impetus to developments in computer science. Through partnership a synergy is created from the strengths and contributions of everyone. When data management strategies are explicitly aligned with the long term goals of synthetic and broad regional research, feedback is ensured to better guide multi-scale database and framework research strategies.

ORGANIZATIONAL STRUCTURE

The development of a hierarchical, networked clearly lies beyond the scope of a single project or institution. The scale and complexity of the task presents a number of challenges for organizing and coordinating the diverse skills and resources within the community of scientists that will be addressing these goals.

We need to envision a vehicle for continued ecoinformatics activities. We see this taking place at several levels –

- Creating a center to coordinate activities, i.e support a consortium
- Participating as distributed laboratories of ecoinformatics
- Developing mechanisms for cross-fertilization of ideas and technology
- Meeting staffing needs by developing mechanisms for professional development

ECOINFORMATICS CONSORTIUM (EIC)

To capitalize on the strengths of individual partnering organizations, we are developing an Eco-Informatics Consortium (EIC) that acts as the steering committee for the larger community. The consortium serves to formalize the ad-hoc partnerships between groups of scientists currently working to address similar questions and provides a mechanism to capitalize on synergism, increase communication and coordination, and accomplish "collective" goals, while at the same time maintaining individual autonomy.

Administrative functions of the EIC would have some sort of physical location, possibly rotating between institutions. However, research activities will predominately be carried out within a virtual environment, supplemented by workshops, conferences or other gatherings. We envision the structure of this entity to be somewhat along the lines of the Dublin Core or the W3 Consortium. The mission of the EIC is to develop the vision for the larger body and mobilize different pieces of the consortium as needed. The EIC structure enables each member of the EIC to concentrate on areas where they have expertise, whether it be information management, computer science, or social issues, yet have these efforts connected to a larger community.

In terms of the larger body, we envision the EIC to be the hub, around which the consortium members are arrayed like spokes on a wheel. Each of these spokes would then act as a local hub for organizations at the subsidiary levels. For example the LTER network office would act as the hub and the LTER sites would be the spokes of this subsidiary wheel. These sites in turn would be hubs for the local community of organizations surrounding their home institution or their domain science.

Depending upon the circumstances, different members of the consortium could be linked to take advantage of specific opportunities as they arise. This will range from a small group of organizations at one of the auxiliary levels mobilizing to develop a software tool to the entire EIC mobilizing to compete for large federal IT grant opportunities. We can think of the functioning of the EIC spokes/hubs as different sets of lights that blink on and off to take advantage of opportunities at a range of levels. This flexibility and balance between small and nimble and large and powerful will be crucial to the success of the EIC.

Information flow through this model of the EIC is critical to the success of its efforts and we envision a bi-directional flow of information, such that solutions can percolate up through the subsidiary wheels to the EIC hub and then be redistributed to the other components of the EIC. This bi-directional flow will allow each member of the consortium to take advantage of all the data, information, and knowledge of the entire EIC.

Distributed Laboratories for Ecoinformatics

While the EIC is expected to exist primarily as an organizational vehicle, we anticipate that it will function to catalyze the creation of many shared infrastructure resources. The most obvious resource will be the network of data sets that can be shared among partners, enabling access to a wider range

of data while at the same time allowing institutions to concentrate their management efforts on those data sets for which they have direct responsibility. Recent work at San Diego Supercomputer Center demonstrates the potential for meta systems - virtual machines that can compile resources from a heterogeneous computing environment, accessing data, memory, processor cycles, and resources from distributed physical computers across a network; network environments for application sharing; and distributed super-computing. Another project of the EIC will be to create collaborative environments that facilitate communication through network technologies such as video-conferencing, white boarding, and application sharing. Finally, we envision the creation of training and educational programs in which the skills and knowledge are disseminated.

Mechanisms for cross-fertilization

The EIC provides a method for organizing and implementing a variety of mechanisms proven to be of value in communicating ideas, technology and innovation. One such mechanism is the organization of topic-focused workshops. NCEAS provides a successful model for how such workshops can be managed. Another is the creation of cross-institutional positions such as the LTER postdoctoral position at San Diego Super Computing Center. Short-term visiting researcher opportunities for personnel involved in information management would also serve to cross-fertilize ideas and solutions. Exchange programs could be developed between sites, between sites and the Network Office, or even between LTER and other partners within the EIC. The tradition of external invitees to the LTER data management meeting has benefited both the LTER program and its guests.

Infrastructure Changes within LTER data management

Expansion of the scale of information management from the local to the network level will carry implications for the organization of IM personnel. First, it is important that information management be well represented in management and executive decision-making. This may require IM staffing changes such as redefinition of existing responsibilities, training for and creation of higher-level positions for IM, or assignment of existing PI-level personnel to IM leadership roles. Second, opportunities need to be created to give information management personnel time via sabbaticals or similar mechanisms for infrastructure-related projects that may involve collaboration with EIC partners, extended travel, etc. Third, there will be a need for increased funding and opportunities for training as information managers struggle to design systems, update technology for intersite data exchange solutions while maintaining existing services and receiving annual dataset additions. Finally, it is likely that information management staff will need to be expanded to accommodate the broader range of activities, particularly in the areas of information and knowledge management, while at the same time maintaining current levels of support in the area of site data collection and management.

CONCLUSIONS

The LTER is a network with a community of scientists well focused on long-term ecological science and with a community of information managers attending to data management. While the complexity of data handling requirements and expectations has increased, strategies are being developed to create and to benefit from a synergy with technology and advances in computer science.

ACKNOWLEDGMENTS

The authors wish to express special thanks to the editors and reviewers for many helpful comments. This work was supported NSF Grants DEB-9980154, DEB-9634135, DEB-9714833,OPP-9632763, DBI-0111544, DBI-9983132, and EIA-0131958

REFERENCES

[1] Baker, KS, Benson B, Henshaw DL, Blodgett D, Porter J, Stafford, SG. 2000.Evolution of a Multi-Site Network Information System: the LTER Information Management Paradigm, BioScience 50(11): 963-978,2000.

[2] Robert J. Robbins 1996. BioInformatics: Essential Infrastructure for Global Biology. Journal of Computational Biology

[3] Michener, W.K., J.W. Brunt, J.J. Helly, T.B. Kirchner and S.G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7(1):330-342.

[4] Jones, M.B., C.Berkley, J. Bojilova, and M. Schildhauer, 2001. Managing Scientific Metadata. IEEE Internet Computing Sep-Oct 2001

[5] Cook, R.B., R.J. Olson, P.Kanciruk, L.A.Hook, 2001. Best practices for preparing ecological data sets to share and archive. Bulletin of the Ecological Society of America 82(2)

[6] Brunt JW. 1999. The LTER network information system: a framework for ecological information management. Pages 435-440 in Aguirre-Bravo C, Franco CR eds North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources; 2-6 Nov 1998; Guadalajara,l Mexico. Fort Collins (CO): US Department of Agriculture, Forest Service, Rocky Mountain Research Station. Proceedings RMRS-P-12.