LTER DataBits
Information Management Newsletter of The Long Term Ecological Research Network

◆ *Feature Articles*

◆ *News Bits*

◆ *Good Reads*

◆ *Frequently Asked Questions*

◆ *Calendar*

Featured in this issue:

According to most of the plenary addresses of the 2000 All Scientists Meeting in Snowbird, Utah, ecological researchers must "think outside the box". Sociologists, modelers, climatologists, paleoecologists, remote sensing specialists, and science administrators all emphasized the need of ecological scientists to become comfortable with multiple disciplines and define questions that are significant in each of these fields simultaneously. Only through such mutual respect and understanding can multiple disciplines be merged into a coherent interdisciplinary research team. In this issue, we pick up this theme and encourage researchers and information technologists to move beyond the traditional role of information management as a data archiving service. We have a unique set of skills to offer ecologists, and those skills are more likely to be used when we define our expertise in relation to pressing ecological questions. The Fall 2000 issue of Databits presents the many ways information managers and technology experts are lending their unique talents to address complex ecological issues and problems today. We have included summaries from information management workshops held at the All Scientists Meeting in the hopes that the readership of this newsletter will envision applications of their own research across the multiple disciplines that comprise ecological science.

◆ Feature Articles

Partnership between Long-term Ecological Research and Information Management: Successes and Challenges

-Barbara Benson, North Temperate Lakes (NTL)

Organizers: Barbara Benson (NTL), Dick Olson (Distributed Active Archive-Oak Ridge National Laboratory, DAAC-ORNL), John Magnuson (NTL)

Since the establishment of the LTER network and partly because of its existence, ecology has expanded from the traditional study of a site or an event by a single individual. The discipline is now much broader, including networks of sites and communities of investigators who use modeling, synthesis, assessments, and long-term data sets. A key factor in this fundamental change is the dramatic increase in the application of computer science to ecology. This workshop generated a productive dialogue among researchers and information managers by examining their mutual partnerships at individual sites and in intersite research. Invited speakers set the stage by presenting an overview of successful partnerships between ecological research and information management (IM).

As information technology continues to evolve and the research agenda broadens, new challenges will need to be met by the partnership. The presentations and discussion attempted to articulate what these areas of growth will be. An objective was to influence the future of partnership between traditional ecological science and IM to support new and innovative projects. We encouraged our audience to translate the discussion into actions, new money, new programs, new emphases.

The workshop was structured as two panel discussions, each introduced by a guest speaker. The first panel focused on the components of successful synergistic partnerships between scientific researchers and information managers. The catalyst speaker was Bill Michener (NSF) and panelists were John Helly (San Diego Supercomputing Center, SDSC), John Magnuson (NTL), and Susan Stafford (Colorado State University). The second panel focused on "big science partnerships" and the new areas of challenge. The catalyst speaker was Bruce Hayden (VCR) and panelists were Peter Arzberger (SDSC), James Brunt (NET), and Jim Gosz (SEV). The discussions included audience participation, with time for generating conclusions and recommendations.

Bill Michener emphasized the need for partnerships between scientists and information managers to support long-term, broad-scale, and synthetic research both at individual sites and for intersite research. He encouraged us to transcend boundaries- to think outside the traditional box along dimensions of time, space, and parameters (Figure 1). Examples of new initiatives to transcend boundaries are NSF's Biocomplexity and NEON programs. Bill maintained that traditional ways of doing ecology do not scale up and that we are undergoing a paradigm shift from a hunter-gather mode of doing research to a harvester mode with integrated, interdisciplinary databases.

Developing successful partnerships may require new skills, new technology, new datasets, new integration methods, and new visualization tools. Such joint efforts between Principal Investigators (PI) and IMs will often result in synergism with the resulting product being greater than sum of its parts. Bill proposed that we can enable successful "synergistic" partnerships if we:

- move from reductionism to integration,
- grow the resource base (dollars),
- develop new ways to access and present information to the community,
- build new tools for instrumenting the environment,
- adopt collaborative analysis and promote teamwork, and
- strengthen the work force of information managers and statisticians.

Incentives for partnerships include:

- new scientific and funding opportunities (e.g. the National Science Foundation's NEON,

Biocomplexity, ITR competitions);

- more venues to receive credit (Ecological Archives, tenure & promotion); and
- directives encouraging partnerships (funding agencies, journals, societies).

Bill concluded by defining ecoinformatics as "a broad interdisciplinary science that incorporates both conceptual and practical tools for the understanding, generation, processing, and propagation of ecological information".

Susan Stafford led off the panel by describing the LTER-NACSE (Northwest Alliance for Computational Science & Engineering) Partnership to promote innovative approaches to data access and analysis. In a vision of the Laboratory for the 21st Century, enhanced access to data, information, and knowledge is provided to a range of customers --- researchers, students, managers, policy makers, and society. Susan emphasized that it is important to know the agendas for each partnership that is created. She contrasted collaboration with NACSE whose agenda was computer-science focused software tool development and usability with collaboration with the National Center for Ecological Analysis and Synthesis (NCEAS) whose agenda is ecology-focused conceptual design and hypothesis formulation. She discussed how the LTER network has created an opportunity by addressing standardized goals with a variety of solutions, rather than imposing strict standards.

John Helly discussed the successful work to build machinery for data publishing that can help promote data sharing and save legacy data. Important developments included ESA's report on the Future of Long-term Ecological Data (FLED), a collaboration to design an ecological data repository by SDSC, ESA, NCEAS and the Bishop Museum, publication of data sets in Ecological Archives, and the release of the 'Caveat Emptor' Ecological Data Repository (CEED, http://ceed.sdsc.edu). He contrasted a variety of electronic publishing models from centralized to decentralized with decentralization being either federated or peer-to-peer. Protocols for publishing range from editorial board peer review to cooperative standards to totally open (Figure 2).

John Magnuson described a successful scientist-IM partnership to support research based on assembling a global lake and river ice phenology database. These data demonstrate a worldwide trend in ice phenology toward later freeze and earlier thaw. The integrated information enabled broader scale research. This project attributes some of its success to the workshop format used. Data were incorporated into a database prior to the workshop and thus were available for initial analyses at the workshop. The researchers negotiated a data access policy. The database is now being released to the public by transferring it to the National Snow and Ice Data Center.

Panel 2 - **"Big Science" Partnerships**

Bruce Hayden discussed successes and challenges and opportunities associated with Big Science Partnerships. He encouraged the audience to continue to think big- really big- in proposing new activities. The successful inclusion of IM in research proposals and the 2900 LTER data sets that are online and available for sharing are evidence of successful partnerships. Bruce pointed out the change in IM from data bankers to information mangers with some IM individuals given "PI" status within a project. Emerging access agreements among scientists illustrate the social aspects of data and information sharing. This environment has led to the growth of new concepts of data and information. We think of data in new ways, and we are doing more because of the agreements and collaboration that lead to the division of IM labor in big science consortia. Finally, NSF has highlighted the value of IM/ecological science partnerships in big science

through new funding programs such as KDI and the forthcoming NEON program; in each case, IM will take a high profile in calls for proposals.

Bruce then summarized challenges and opportunities that are the drivers for partnerships. The science consortia and consortia of consortia (e.g., LTER, ILTER, NEON) are trends that will "drive" us together. However, the challenge will be to have a better structure for building consortia and to build on the successful networks. On the technology side, the smart sensors (field instruments with IP addresses) will become more standard; for example, "Sci-Net" (over the horizon communications) technologies will provide networks using a wireless approach. Challenges to the IM community include the archiving or curation of data (SDSC provides an example with its Storage Resource Broker), metadata standards (FGDC, ISO), metadata crawlers customized for scientists, and science "cookies" to harvest selected information. Finally, nanotechnology is beginning to offer more work for less mass and energy; these technologies will be great both for the environment and information transfer.

James Brunt outlined the evolution in levels of understanding and action in IM over three decades. The 1980's were primarily devoted to developing data management methods, the 1990's to developing the delivery of network information, and the 2000's to building the data, information, and knowledge infrastructure for advancing ecological science. Ecoinformatics infrastructure components (Figure 3) include networked data storage, tools and procedures for facilitating data integration and synthesis, and resources to promote and support synthetic research.

Peter Arzberger asked scientists to volunteer to serve as "drivers," collaborators, and dreamers to help test and develop an integrated advanced computational infrastructure. He emphasized that today's partnerships are building the infrastructure for tomorrow and there is need for applications and users for this process. This is a time of great opportunity as digital information explodes, bandwidth increases, wireless technology permeates, and high performance computing/information technology pervade science and society. Peter emphasized how partnerships will need to use integrative approaches to address such issues as sustainability science. He described the progress under the NSF funded NPACI (National Partnership for Advanced Computational Infrastructure) program to build and provide an advanced infrastructure to do science and make information accessible to others.

Jim Gosz emphasized the huge potential to use nanotechnology at small scales. There are opportunities to do big science both toward the atom and toward the globe. Jim suggested that partnerships with those groups doing nano-scale work exist because they have money but need ecology. We need ways to make big steps to "catch-up."

Summary of comments

The differences between the two panels were contrasted: the successful partnerships panel involved traditional data issues while the future partnership panel minimized the data issues and emphasized information management. Under this second view, scientists need to realize that information technology is a crucial driver. While data ownership may still be an issue in the future scenarios, data must be more freely contributed - possibly with a top down approach to get people to submit data.

The discussion emphasized an urgency to participate in the big science and use new technology to address science issues. While NSF is pushing high technology, there is a realization that there must be ecologists involved. Based on past experience, partnerships are a way to accelerate the rate of using new technology. An example is how the oceanography community responded to the NSF funding of ocean voyages: they

developed a strategy of research with more sharing of resources and data. Within this context, there is the issue of giving credit to people doing data work and how data sharing and group publications will be included into tenure review, etc. These may be less of an issue with the increasing automation of data collection.

The lessons learned by existing partnerships need to be shared with others, both successes and failures. One experience shared by several speakers was that training scientists in IM techniques was generally more successful than using pure computer scientists. Also when the IM group was isolated from the science, many groups experienced difficulties. Standardization is a big topic for sharing data, and sometimes a top-down approach is needed. However, interoperability is another approach that can be used to combine new data or to bring in older data. It is important to have scientists and data folks understand their roles in a partnership, even to the point of formalizing organizational structure for communication and making sure there are good communications. Finally, suggestions were made to draw on studies of partnership in other disciplines to help identify the "chemistry" needed for a successful partnership. We need to document the benefits of collaboration and help new groups to move into this mode of

Recommendations:

1. The Top-Down Approach. Lobby the academic communities to give recognition (tenure, promotions) for teams addressing big science questions and lobby funding agencies to provide resources to conduct big science (group proposals, longer funding periods, new measures of performance).
2. The Bottom-Up Approach. Provide incentives for scientist-IM partnerships such as open-literature publishing of data, Internet distribution of data, IM training for ecologists, generic and customized IM tools, and funding for synthesis workshops.
3. Encourage data integration and synthesis across space and time scales for multiple parameters to address broad-scale science questions (and encourage a strong IM component).
4. Promote the use of technologies such as eddy-correlation flux towers, smart sensors, networks, Nano Smart Dust, etc. to address key ecological questions and promote the need for IM to support the high data volumes generated by these new technologies.
5. Make sure that ecologists are active in the development of new information technologies and interpretation of resulting data, that is, encourage scientists to be leaders in the emerging big science fields.

Advanced Communication and Networking: Opportunities and Challenges for LTER/ILTER

- William Y.B. Chang, National Science Foundation

Purpose:

1. To identify opportunities, challenges, and applications of advanced communication and networking for researching and teaching in LTER/ILTER communities; and
2. To integrate these cutting edge tools to support end-to-end applications in ecological and biodiversity monitoring and research.

Convenor: Bill Chang
Speakers: David Hughes, Old Colorado City Communication, on Wireless Data Acquisition and Communication
John Jamison, STAR-TAP/Juniper, on Regional and National Networking vBNS/I2 and Beyond
Reagon Moore, SDSC, on Integrated Data Management and Analysis
Tony Fountain, SDSC, on Environmental Application of Advanced Computing Infrastructure
John Vande Castle, NET, on Challenges and Bottlenecks

Outcomes:
Emerging technologies in communication and networking and the development in high performance computing are critical to continuing success in ecology and to addressing complex problems in biology. The purpose of this workshop is to bring together a group of leaders in this area to discuss the current level of development in this area, to demonstrate new applications using these technologies and to address challenges and bottlenecks in using these developments in LTER and ILTER communities.

The major points presented in the meeting include:

- Wireless technologies can play important roles, especially in areas with limited or outdated telecommunication infrastructures, as shown in the application of wireless communications in Ulan Bator, Mongolia. Success in such applications was also shown in the canopy study at the Luquillo Experimental Forest Site.
- Extensive high performance networks, such as NSF's vBNS, Internet 2, and STAR -TAP (in Chicago, inter-connecting with ten networks in Asia, Europe, Russia, and Israel) are available and can be useful in facilitating scientific data exchange, model simulation, communication, and teaching.
- New database and management systems, such as the Storage Resource Broker (http://www.sdsc.edu) technology, can improve the integration of heterogeneous datasets in scientific research.
- New environmental applications using these technologies can provide new insights into environmental and ecological problems. We must simultaneously overcome the following hurdles to address ecological problems: (1) large spatial extents and fine detail; (2) integration across disciplines; and (3) linking model simulations with data.
- LTER research faces many challenges and bottlenecks to integrating these emerging technologies, iincluding knowledge divides, information divides, limited availability of facilities, and lack of communications infrastructures.

Technology Focus: Environmental Applications of Advanced Computing Infrastructure

- Tony Fountain, San Diego Supercomputing Center

As research goals for the LTER Network shift to more inter-site and regionalization studies, the role of communications and computing infrastructure becomes more significant. The end goal is a seamless integration of tools that supports end-to-end applications in ecological monitoring and research (i.e., data collection, communication, management, analysis, application). Significant progress has been achieved towards this goal as evidenced by the applications and infrastructure components described in this workshop.

In this presentation, I identified five areas of the SDSC/NPACI high performance computing infrastructure that have immediate application to research activities of the LTER Network:
1. Data and collections management.
2. Wireless communications.
3. Hi-performance computing, modeling, simulations.
4. Scientific visualizations.
5. Data analysis, mining, decision support.

In each of these areas, SDSC and the LTER Network are collaborating on specific applications. A few of these are described briefly in the sections below. Interested readers are encouraged to contact the principals through the URL's provided at the end (or contact me).

Data and collections management
Data, or collections, management is a pervasive issue in LTER Network research activities. SDSC and the LTER Network have two ongoing collaborations in scientific collections management; both employ the SRB technology described in Reagan Moore's presentation.
1. KDI-SRB Project - This is an NSF funded project in collaboration with the National Center for Ecological Analysis and Synthesis (NCEAS). The primary goal is to use the SRB technology to federate LTER Network data sets. In this case the SRB technology is being used to logically combine selected heterogeneous databases of LTER sites into a single data resource. Five LTER sites have been identified to participate in the initial deployment.
2. LTER-hyper-SRB - This is a small prototype project that combines the SRB and the SDSC high performance storage system (HPSS) to create a collection management system for LTER hyperspectral data. In this case the SRB is being used as a collections management system for a single large data archive. This system is currently in beta production mode.

Wireless communications
The National Laboratory for Applied Network Research (NLANR) has as its primary goal to provide technical, engineering, and traffic analysis support of NSF High Performance Connections sites and HPNSP (high-performance network service providers) such as the NSF/MCI very high performance Backbone Network Service (vBNS). Under the direction of Hans-Werner Braun, the NLANR group at SDSC has developed a wireless backbone for San Diego County and they have been funded recently to extend this to encompass broader geographical areas and diverse applications. Initial applications include the collection of astronomical data from observatories (Palomar and Laguna), seismic data from east county deserts (geophysical measurements), as well as meteorological data.

Hi-performance computing, modeling, simulations
Computer simulations are critical tools in ecological research, and their role will increase as communications and computing resources improve to support techniques such as real-time data assimilation. The high-performance computers at SDSC provide the computing complement to the network and data management infrastructure. There are three trends in large-scale ecological simulations:
1. Increased scales and finer resolutions (time and space)
2. Integration across disciplines (e.g., atmosphere, biosphere, ocean)
3. Integration of simulation and sensor data (e.g., data assimilation)

The NPACI IBM Blue Horizon supercomputer at SDSC is the most powerful machine available to the scientific community. Recent experiments by Joe Eastman (LTER Network Office) with a coupled atmosphere-ecosystem model have demonstrated the significant role that these high-performance

resources can play. He went from 30x30x20 gridpoints at 50km grid spacing horizontally, and 250-2000m vertically to 150x150x30 gridpints at 10km grid spacing, with 12.5m to 1500m in the vertically. The timestep was also decreased from 120s to 30s. The results from these scaling experiments were presented at another workshop at LTER ASM 2000.

Scientific Visualizations

The growth of data collections (field measurements, remote sensing, computer simulations) and the need to integrate data across geographical areas and scientific disciplines necessitates new methods for analysis and visualization. The Scientific Visualization Group at SDSC has been collaborating with LTER Network scientists to apply high-performance and non-traditional visualization techniques to LTER Network data. Nicole Bordes, SDSC Visualization Group, has conducted a number of tutorial and consulting sessions for LTER Network visualization projects. A sample of these is viewable on the SDSC Visualization Group web site (see URLs below). Mike Bailey of SDSC performs a second type of scientific visualization that has application to LTER Network scientists - ChromaDepth 3D images (with 3D glasses) and 3D physical models constructed on SDSC's Laminated Object Manufacturing (LOM) machine. The finished LOM models have the appearance of sculpted wood and provide unique perspectives on the modeled landscape. High-performance computer visualizations and 3D models provide useful tools for understanding LTER Network data and SDSC and the LTER Network are collaborating to find the best ways to apply these tools to LTER scientific activities.

Data Analysis, Mining, and Decision Support

As data collection rates increase, and data archives explode, it becomes more critical to find efficient methods for deriving knowledge from these sources. The existence of large data warehouses of environmental data challenges the traditional paradigm of hypothesis-driven data collection. Increasingly, the need to analyze large archives for the scientific activities of exploration, explanation, and prediction, requires new analytical tools. The fields of data mining and knowledge discovery provide tools and techniques for addressing these issues in commercial applications. Extending these methods to environmental data and issues (e.g., geospatial) requires new research and development efforts. SDSC has initiated several projects in this area, and these include the development of new data mining algorithms as well as the application of commercial off the shelf (COTS) products. In particular, David Stockwell and the Environmental Informatics group at SDSC have developed an implementation of the K2 algorithm for the induction of Bayesian networks, and this is being used in biodiversity studies. An NSF-funded study of land use patterns and ecosystem function in Mongolia is also underway. In addition, members of SDSC staff are currently investigating the application of commercial data mining tools and developing the computational infrastructure needed to make these tools useful to LTER researchers. Stuart Gage's lab at Michigan State University (KBS LTER) has been active in this area, and they are currently collaborating with SDSC staff to create elements of a computational grid for performing ecological modeling and data mining.

Conclusion

As mentioned in the introduction, the overall goal is to tie these technologies together for an end-to-end solution: from sensor, to simulation, to visualization... to decision support --- from acquisition to application. These projects are concrete first steps towards this goal. The next steps are to increase functionality and robustness of the existing components and then close the loop by applying these techniques to specific issues in ecological science.

Links
San Diego Supercomputer Center
National Partnership for Advanced Computational Infrastructure

---

Ecological Informatics: Innovative Tools and Technologies

- Hap Garrett, PIE

The workshop entitled "Ecological Informatics: Innovative Tools and Technologies" was organized by John Porter (VCR) and Hap Garritt (PIE) with the intent of providing introductions to new technologies that could provide aid in the input, management and analysis of ecological information resources. The workshop was attended by 46 participants representing 20 LTER sites and 7 other institutions and agencies. The workshop began with presentations by: James Brunt (A Knowledge Network for Biocomplexity: Building and evaluating a metadata-based framework for integrating heterogeneous scientific data); John Porter (Network Collaboration Tools and Technologies); Paul Hanson (Wireless Networking Using Spread Spectrum Technologies); Mark Losleben and Bethaney Swain (Providing Real-Time Video Data-the Tundracam (http:// tundracam.colorado.edu) and Hap Garritt (Microdataloggers and Beyond). Presentations are available for viewing at http://www.vcrlter.virginia.edu/tools2k.

The presentations were followed by discussions of other technologies ranging from bar codes for sample collection; voice recognition software (see John Anderson's review of Dragon Naturally Speaking); voice activated modems for data downloading and radar and acoustic imaging. There was a general consensus concerning the need for a forum for sharing researchers' experiences with new technologies. These shared experiences would help save researchers time in determining which new technologies were worthwhile to incorporate into their research.

Next steps for the group were identified:
(1) set up a List Serve, creating a WWW site of the Ecological Informatics workshop presentations (http:// www.vcrlter.virginia.edu/tools2k);
(2) start a chat room to share experiences with new technologies; and
(3) set up an electronic clearinghouse with a form on the WWW for personal reviews of equipment use.

GIS on the Internet and LTER: a Frontier for Research, Applications, and Solutions

- Ned Gardiner, CWT

This workshop convened to discuss the ways LTER sites have used and can use the internet to serve geographic information systems (GIS) data and functionality to diverse users: investigators, students, collaborators, policy makers, and citizens. LTER has evolved with the internet; in the 20 years since its inception, the internet has grown from a US security tool to a common household amenity. Over that 20 years, information managers have led other LTER investigators in using the net to link sites and in reaching

out to a broader community of researchers and citizens. This workshop enhanced our dialog about providing geospatial data and content over the World Wide Web. This is a frontier because it will require (1) interactions across disciplines and (2) applications designed to meet the specific needs of research projects designed within disciplines other than computer science and geography.

The World Wide Web is a powerful way to present complex GIS data to users with varying levels of experience and understanding of GIS concepts. Often, people associate the term, "GIS", with a software vendor, a data model, or a set of techniques. In reality, GIS is a very broad subject area comprising the personnel, hardware, software, and support structure needed to store, analyze, and retrieve spatial data. GIS has become commonplace in the research environment, but it remains a complex subject matter. GIS practitioners must be comfortable with concepts in computer science, geography, networking, programming, and all the subject areas touched by the data a GIS supports- geology, climatology, ecology, geomorphology, hydrology, plant physiology, and soil science, to name a few that often come into play in ecological research. Internet applications an attractive potential: simple and intuitive presentations of complex geospatial data.

But that potential requires diligent attention by qualified specialists if is to be realized. The ecological and biological research communities can participate in developing internet GIS technologies by identifying their common needs. With these themes in mind, the objectives of this workshop were to:

1. Introduce internet GIS technologies to information managers and researchers.
2. Demonstrate examples of how internet GIS has been useful in research, visualization, and communication.
3. Provide a forum to exchange ideas about what has and has not worked. This forum was aimed at researchers and technicians. We hoped researchers would get inspired about applications and that GIS technicians would exchange successes and failures and explore the future direction internet GIS applications should take to serve LTER network needs.
4. A final objective, then, was to explore how we can use various technological approaches to solving research, visualization, and communication problems within the LTER community.

Ned Gardiner: A Frontier for Communication, Visualization, and Research

Ned Gardiner (CWT) introduced the workshop with examples from Coweeta. Coweeta's GIS web site has grown from a listing of data sets and maps to an integral data management tool. In 1998, Coweeta began to integrate spatial and non-spatial data by developing metadata and data dictionaries (Figure 1). This work continues as a collaboration among each scientist (co-investigators, visiting researchers, graduate students, summer students, etc.), the site manager (Brian Kloeppel), the data manager (Ron Rouhani), and the GIS manager (Ned Gardiner). Gardiner presented three examples of how database documentation and integration have enhanced visualization, research, and database management. The first example involved archiving and serving data via the web site. The success of this effort has been tempered by a general lack of sufficient metadata to serve these data sets via national programs such as the National Spatial Database Infrastructure (NSDI). The second example used Arcview Internet Map Server (IMS 1.0) on a Windows NT server running Microsoft Internet Information Server (IIS). The arcview extension uses Java applets to provide live mapping capabilities to any user with an internet web browser. This technology was illustrated with a site selection example. Coweeta stream researchers selected 8 sites in the Little Tennessee and French Broad river basins to examine expected development impacts on stream ecosystems. The Coweeta research team works at several university campuses across the country, so the internet tool facilitated communication and visualization. Drawbacks included transfer bottlenecks across the internet and the

speed of the machine used as a server. The third example related of fish assemblage data to modeled sediment inputs in a large river basin. Fish collections, model results, and spatial attributes of sampling sites were stored in an mSQL (Hughes Technologies) database on a Solaris platform. Data from these sources were summarized, integrated, and analyzed using Perl. The Perl script accessed the mSQL database and ran a SAS program.

Brian Kloeppel, Ned Gardiner, and Ron Rouhani. Interactive Mapping for Research Project Management

The last example above was designed like Brian Kloeppel's (Coweeta) project. He needed a tool to keep track of research projects within the Coweeta Basin and larger study region. The examples used html, perl, mini-SQL (Hughes Technologies), and Arcview IMS 1.0. The prototype he described allows users to discover what research projects are occurring in the vicinity of a particular location within the Coweeta Basin or in the larger study region. This site management tool allows the user to access and/or download images, GIS data, metadata, and data sets. Given the relational structure among tables in the Coweeta database (Figure 2), queries may be spatially explicit or based on theme attributes, researcher, sampling dates, or other attributes stored in each data set's metadata. This links data collected for specific scientific applications to geospatial data. These scripting and database tools were already used in managing the data collected by field researches, so this was a quick way to integrate the spatial and thematic data at Coweeta. Information managers did not have to use proprietary spatial database integration tools.

Theresa Valentine: GIS on the Internet at the HJ Andrews LTER Site

Theresa Valentine (H.J. Andrews) presented an overview of some of the technologies offered by ESRI to do internet mapping. She collaborated with NACSE to produce a Java/Arc IMS 2.0 application to serve and manipulate data from the H.J. Andrews Experimental Forest. Users can turn themes on and off, change the scale of the view, and execute spatial and non-spatial queries using a common internet browser. Her application underscored the importance of (1) detailed metadata and (2) adopting frequently-changing technologies that allow one to link spatial data sets with non-geospatial data. Her project required team work, computational architectures capable of serving complex software over the internet, and working with non-GIS users to develop applications they could understand. These concepts will be crucial for other sites to succeed in similar ventures.

Peter McCartney and Matthew Luck: Internet Delivery of Spatial Data at CAP LTER

Peter and Matt pointed out that there is vanishingly little difference between "spatial" data and "non-spatial" data in the ecological sciences. Whether or not an investigator is a geographer collecting data specifically for geospatial data processing, all ecological data pertains to a specific site at a specific time. Hence, information managers should be trained to think about how these implicitly spatial themes can be stored and retrieved in a GIS. The availability of spatial and ecological data via the internet require urgent attention to the details of how to link heterogeneous data sources. They presented an intuitive spatial data search and retrieval system. Their application facilitates the researcher's own view of his or her data, rather than imposing a "GIS" view of those same data. They used ESRI's MapObjects software with Visual Basic.

Development of internet mapping at CAP LTER has focused on several goals: seamless access to metadata and data; close integration between spatial and non-spatial resources to allow queries to reference both spatial and categorical options; online visualization and processing of remote data resources prior to

download. A three-tiered system includes web services based on MS NT, a data archive consisting of MS SQL Server, ESRI Spatial Database Engine, and some file-based datasets; and a spatial processing application built with ESRI MapObjects and MS Visual Basic. An early application developed for CAP was a map-based query tool for browsing data from the bird survey project (http://caplter.asu.edu/po12). A more advanced application is now being developed to add spatial enhancements to the existing CAP online data catalog (http://caplter.asu.edu/data). Portions completed thus far allow users to select spatial data layers from the data catalog to be viewed with full zoom and pan capability through the map utility. Users can then define a clip area based on the view window or several predefined boundaries and select a desired output projection. The clipped and reprojected layer is then zipped and presented as a temporary link via the standard CAP LTER download page.

The approach taken with this utility figures prominently in CAP's plan for replacing direct FTP access to data archives with a three-tier server based system that brokers requests for data. This hides internal details of file storage from the general public while at the same time permitting the site to supply advanced processing features as part of the data access service. Presently these features are limited to viewing, subsetting, and reprojecting data but will eventually be extended to include more advanced operations such as data summarization, analysis and model parameterization.

Chaitan Baru: XML-mediated GIS

Chaitan presented a series of technologies the Data Intensive Computing Environment group at San Diego Supercomputing Center has developed that use XML as the mediation language for internet database and GIS applications. XML is supported by industrial software providers. Chaitan contrasted centralized databases, in which all data are tightly linked, with distributed databases that require a mediator to pull data and information together on the fly. The latter, loosely-integrated scheme represents much of the data in the LTER network, so we are likely to be increasingly concerned with XML as a mediation language for internet-based applications, such as GIS operations. ESRI's model for using XML to distribute GIS data via the internet may be found at:

http://arconline.esri.com/arconline/index.cfm?PID=6

Action Items

Each talk offered a unique perspective on a set of issues that most sites share in common. For example, how do we ensure that legacy data will be usable as we integrate our spatial and non-spatial data? How do we need to improve the information management process so that data are integrated as they are archived, rather than *post facto*? What applications are needed? When should one use a full-blown application or the client-server paradigm?

The group also discussed forming a GIS working group within the LTER community and developing cross-site and intra-site research proposals that would help push internet mapping technologies in directions appropriate to ecological research.

GIS Working Group

First, we agreed that there were many common research and information management needs identified among the talks. We saw equally many technological implementations to meet those needs. It would be useful to explicitly identify the needs that each site has and develop a working group that could provide feedback to the technology community to build tools to suit those needs. Sites should not be left to their own budgets and expertise to individually develop software solutions. The working group should address the following issues:

1. How to support raster data using internet GIS technology.
2. Open GIS interoperability. Do existing web mapping testbed (WMT of the OpenGIS consortium) levels meet expected needs?
3. Provide feedback to ESRI's internet mapping group.
4. Identify extensions to existing technologies particularly suited to LTER.
5. Examine data exchanges standards. What are the advantages of XML options such as vector markup language (VML) and geographic markup language (GML)? How do these compare with ESRI and ERDAS formats, in a practical sense?
6. Participate with other partners outside of the LTER network doing similar work.
7. Identify the base technology that exists in the LTER network and beyond. Further, what technology is needed for the technological innovations the work group identifies as most promising for site and intersite research?
8. Finally, provide practical solutions and advice to sites who are integrating spatial and thematic data. This effort must seemlessly link metadata and raw data.

Research Testbeds

It would be useful to develop a cross-site research project that utilizes internet technologies. Such projects would solve geospatial analysis needs common to many sites and build important linkages with research groups not currently using these internet technology to their advantage. The following groups, represented by work group sessions at the 2000 All Scientists Meeting, were suggested as possible collaborators:

Land Transformation
Land Use Change
Biological Legacies

Within-site research might also contribute to cross-site research, for example the work groups dealing with:

Regionalization (http://www.cevl.msu.edu/regionalization)
Hierarchical patch dynamics (CAP)

Workshop participants will continue to work on the themes presented above, so feel free to contact us with your comments or interests.

News Bits

Into the Fray: Observations by a New LTER Information Manager

- Wade Sheldon, GCE

When I was asked to step in and write the data management section of an LTER proposal over a year ago, I immediately envisioned myself sitting comfortably in front of a gleaming computer coding colorful database applications all day. Now, as I complete my first month as data manager of that LTER site, my view is somewhat different. The gleaming computer is still an order making its way through the purchasing system, the colorful database applications are hiding behind a menacing row of programming books on my shelf, and I realize that there is a whole lot more to managing ecological data than just numbers and programming.

In some ways, working at a new LTER site has probably made my start in data management easier. I have the chance to research and consider various organizational and technological approaches without any constraints imposed by legacy systems or practices. This freedom should allow me to take full advantage of best practices and emerging standards within the LTER (e.g. XML -based approaches for metadata exchange and archiving data, web-centric interfaces for data entry, retrieval, and analysis). I also have the opportunity to assemble an information system using up-to-date hardware and software components, without major concerns about backward compatibility with other components. On the other hand, I often feel overwhelmed as I study the bewildering array of information technology options available today. I also feel a tremendous sense of responsibility, realizing that paths I choose today will affect our project for years to come. Budgetary constraints and the great diversity of computer platforms, software, and network technologies already in use on our project further complicate this process.

As I sort through these issues I am increasingly grateful for the wealth of ecological information management knowledge available online and in print, largely due to the efforts of the LTER data management community. The review papers and site overviews presented in Data and Information Management in the Ecological Sciences: A Resource Guide (Michener et al., 1998), and updated in Ecological Data - Design, Management and Processing (Michener and Brunt, 2000) are invaluable resources, both as guidelines and entry points into the eco-informatics literature. The site overview case studies are particularly informative, providing many practical examples of the principles, problems, and solutions described in the more theoretical chapters. I have also learned a lot from the less formal online resources as well, such as the LTER software surveys, reports and minutes from past data management meetings, and DMAN mailing list discussions. Learning from this combined experience, gained over two decades, has increased my decision-making confidence immensely.

I would characterize my initiation into LTER data management as very smooth overall, but there is certainly room for improvement in the process. One early frustration was the lack of centralized information on network-level responsibilities and resources for data managers. Karen Baker recently began to address this issue by creating the Information Managers Guide web page, but I would strongly encourage the network office to create an orientation packet for new data managers, containing a list of resources and services provided within the LTER Network and the contact person for more information. Other worthwhile additions to the packet or IM Guide would include the following: a calendar of submission deadlines for information management activities, e.g. survey input and publication deadlines for items such as Databits articles; a repository or discussion board for learning about information technology techniques; and an expert referral list to encourage learning and technology exchange within the LTER IM community. Personnel turnover within our ranks will probably remain a major issue in today's Internet economy, so more effort will have to be devoted to training and information transfer in the coming years.

This has been an exciting first month, and I look forward to the many opportunities and challenges that lay ahead as we begin our LTER program.

References

[Ecological Data: Design, Management, and Processing. 2000](). William K. Michener and James W. Brunt - Editors. Blackwell Science, Methods in Ecology Series. 192 pages.

[Data and Information Management in the Ecological Sciences: A Resource Guide.]() 1998. William K. Michener, John H. Porter, and Susan G. Stafford - Editors.

Online Real Time Weather Data at Konza Prairie Biological Station

- Brent Brock, KNZ

Konza Prairie Biological Station recently added real time weather data reporting capability to our Web site (http://www.konza.ksu.edu/). This system supplies graphical displays of 5 minute, hourly, and daily interval weather data to an Internet browser. Following is a description of the system's configuration.

Hardware: The system uses hardware and software available from Campbell Scientific and runs on a Pentium 160 MHz computer running Windows NT 4.0 Workstation configured as a Web server running Microsoft Internet Information Server. This workstation is connected via T1 to the Internet. The meteorological station records data from a typical sensor array (LTER level 2 plus soil temperature) to a Campbell CR10 datalogger. Short haul modems are used to create a dedicated connection between the datalogger and the NT box. Details of this configuration can be found at: ftp://ftp.campbellsci.com/pub/outgoing/lit/srm5a.pdf. These modems were ideal for our application because we were able to retask an existing phone line previously used for traditional telephone modem communications. Short haul modems can transmit data up to 6.2 miles at 9600 baud. However, for sites without existing wire, spread spectrum radios may provide a cost competitive alternative.

Sofware: Data transfer is automated using the scheduling capabilities in Campbell's PC208W software. To prevent accidental shutdown of the software, PC208W is run as a service on the NT workstation following Application Note 3C-O (ftp://ftp.campbellsci.com/pub/outgoing/apnotes/208serv.pdf). We store 5-minute interval data in final storage area 2 on the CR10 datalogger and download all data (5 minute, hourly, and daily) every 5 minutes. To save disk space, the 5-minute data is deleted from the workstation each night.

The heart of this system is Campbell's RTDM software package. RTDM automatically updates graphs and tables at user defined intervals or whenever new data is detected. The updated graphs are converted to GIF images that are served on a standard Web page. As graphs are updated, new images overwrite the old ones so the Web content is automatically updated. As with PC208W, we run RTDM as a service to prevent accidental shutdown. RTDM is fairly easy to use and surprisingly flexible. The software allows defining new data items using expressions, which is very useful for making unit conversions or more complex calculations like heat index. RTDM also lets the user set up alarm conditions and execute a command when an alarm

condition is met. We used this feature to greatly speed response time to hardware problems by setting up alarm conditions for each sensor that trigger when out of range values are recorded. When triggered, the alarm conditions execute a $10 shareware utility called MailSend (http://www.radiks.net/jimbo/share.html) to send an email message to the met station administrators alerting them of a possible problem. This message directs administrators to a special administrative Web site containing RTDM-generated sensor data and range flags to help them diagnose the problem.

Cost: Total cost of this system excluding PC workstation, met station/datalogger, and PC20W software is $759. Of this, $432 was for hardware for the dedicated data link and $326 was for RTDM and MailSend software. As stated previously, we used an existing phone cable so our costs do not include materials and labor for installing a line.

Problems and Lessons Learned: Our system has only been operating for 2 months so long term reliability is not known. However, the short haul modems were easy to install and have operated flawlessly. Overall the RTDM software has been reliable and fairly easy to use. Like most software, the help files are a too vague and Campbell tech support had to be contacted for help with some poorly documented controls. Campbell support was also needed to resolve a problem with a malfunctioning heat index equation but they were able to provide a revised, and functioning, equation within a week. On two occasions RTDM stopped updating image files for unknown reasons. Simply stopping and restarting the service fixed the problem and we added automatic scheduled service restarts to increase reliability. Finally, since the PC workstation is installed at the field site, reliable remote administration of the workstation is essential. We use remote control capabilities built in to Novell Netware 5 but software such as PC Anywhere should also work.

Summary: The real time weather system at Konza Prairie Biological Station was relatively inexpensive and simple to install. It was operational within 1 week of delivery. An immediate benefit has been a substantial improvement in QA/QC of weather data collection. The Web site has also proven to be a useful tool for planning prescribed burning, the Schoolyard LTER Program, and researchers planning field activities. Development is underway to calculate and display daily grass reference ET values which will be useful to several ongoing projects. We are also exploring the possibility of adding instrumentation to calculate real time actual ET at Konza Prairie Biological Station.

Bill Michener Joins LTER Network Office

- Robert Waide, NET

I am pleased to announce that Bill Michener has joined the LTER Network Office. Bill's principal responsibilities will include working with the Organization for Biological Field Stations to upgrade their information management capabilities and to strengthen the relationship between that organization and LTER. We anticipate that Bill will become involved in a number of other LTER information management activities. We are very pleased that Bill has decided to join us, and we look forward to working with him.

Henry Gholz, New LTER Program Officer at NSF

- Ned Gardiner, CWT

The following is a brief message from Henry Gholz, the new Long Term Ecological Research (LTER) program officer at the National Science Foundation. Among his highest priorities for the coming year will be 9 site reviews. Below, he comments on the role of information management within LTER and in the site review process.

> As all the sites know, I[nformation] M[anagement] has become an explicit criterion in the LTER site review process, which reflects its importance in the program. This is a long way from simply giving data on a disc to a secretary! In fact, as you note, it's getting to be quite a ways beyond "simply" posting datasets on a site-specific web site (although that is still an imperfect art in LTER). Some sites/IMs are pushing the envelope in this field, and it is fair to say that the LTER program is one of the leaders in dealing with the management and use of large, complex, multi-site datasets. We can't lose sight of the fact that posting well-documented, clear, complete and correct datasets on easily accessible web sites is still the motivating factor, and that we have some ways to go on that. But, the development of new approaches and innovations is clearly required in this rapidly evolving area and I hope that LTER IMs can continue to be among the pioneers.

◆ *Good Reads*

Good Reads: Ecological Data: Design, Management and Processing

- John Briggs (CAP/KNZ)

Good Reads: Michener, W.K., and J.W. Brunt *eds*. 2000. Ecological Data: Design, Management and Processing. Blackwell Science. 180 pp.

This review is very biased. I reviewed this book for the publisher in an earlier version. I liked it then and I still like it. In fact, this book will be a focus of a new course that I am developing next semester (Spring 2001) concerning data management and analysis in ecological science. The editors and the authors have done an excellent job in putting their years of experience into a compact source of knowledge. This is a book every scientist who is associated with a long-term (> 2 years) project should be aware of. More importantly, any person responsible for managing the data of a long-term project should keep this book around them all the time! For most LTER information managers who have been around for a while, attended the annual LTER information manager meetings or who have read most of the material associated with those meetings, you will recognize most of the information presented in this book. If you are in that crowd, don't expect to learn something remarkable and new in this book. However, for the first time you will have a publication that "puts" it all together from thinking about the generation of hypotheses, the nuts and bolts of scientific data bases including archiving to the transformation of numbers into useful information and knowledge. In addition, by reading this book it should help you remind yourself just how important you are to the project. A must-read and have for every LTER information manager.

◆ *FAQs: Frequently Asked Questions*

Where can one find a quick check on computer terms, such as the definition of 'petabyte' or the meaning of 'gif'?

- Karen Baker (PAL)

Try the original 'free on-line dictionary of computing' (foldoc) with more than 13,000 terms and growing: http://www.foldoc.org This is a UK site with a mirror site: http://www.instantweb.com/foldoc maintained by a web hosting company.

◆ *Calendar*

*01 Aug xx-xx. LTER IM Annual Meeting; Madison, WI
*01 Aug 05-09. Ecological Society of America Meeting; Madison, WI
*01 Oct 16-19 KDI Annual Meeting, SDSC. Report by Karen Baker.

NCEAS, SDSC, LTER and Texas Tech partners recently held an annual meeting (16-19 October) for the Knowledge Network for Biocomplexity (KNB) project (NSF/KDI 99-02) at the San Diego Supercomputer Center. There were a number of working group reports: the NCEAS metadata team is developing the Ecological Metadata Standard EML, the cross-site metadata catalog Metacat, and the client software Morpho; the SDSC data storage team is focusing on development of the storage resource broker SRB with MCAT; the LTER Network Office SRB implementation team is working on SRB installation with LTER site partners and interfacing with data managers Louise Johnson at Cedar Creek, Brent Brock at Konza Prairie, Jim Laundre at Arctic Tundra, Garrett Ponciroli at Kellogg Biological Station, and Mike Hartman at Niwot Ridge; the SDSC hypothesis modeling engine team is working with Baysean modeling for biodiversity data; and the biocomplexity research team is focusing on the relationship between biodiversity and primary productivity. The meeting highlighted project progress and provided a forum for discussion of ongoing and future integration. To stay current with potential interfaces and impacts for LTER sites, be sure to familiarize yourself with the KNB website:

    http://www.nceas.ucsb.edu/kdi

where information is available including a project handout and a powerpoint presentation. This project's accommodation of data heterogeneity and its plan to enable advanced services for data integration makes

their mission pertinent for any site with interoperability requirements.