**Strategic Plan for Cyberinfrastructure in the LTER Network**

## Table of Contents

## Executive Summary

The National Science Foundation (NSF) Long Term Ecological Research (LTER) network has articulated new visions for research that seeks understanding of human-natural systems through advances in collaborative, synthetic social-ecological science at multiple scales. This report defines improvements in cyberinfrastructure (CI) that are necessary to facilitate this research and to support other ongoing LTER research activities. As defined by Atkins et al. (2003), “In scientific usage, cyberinfrastructure is a technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge.” In our context cyberinfrastructure embodies both the people and the technologies that allow collaborative activities and technological solutions for data collection, discovery, access, integration and analysis across disciplinary and scale boundaries.

To identify cyberinfrastructure challenges and consider potential solutions, LTER CI planners convened a diverse group of information technology (IT) professionals from science and technology centers, large IT development projects, and national observatory initiatives in a series of meetings that addressed (1) Multi-site Experiments, (2) Data Integration, (3) Modeling, and (4) System Architecture and Human Resources. These groups identified areas where improvements in cyberinfrastructure are necessary, including data acquisition, management, and curation; data discovery, access, and integration; modeling, analysis, and synthesis; and large-scale collaboration. Crosscutting issues that span cyberinfrastructure improvement areas include development of a service-oriented architecture on which to build collaborative environments, strategic CI partnerships, programs for workforce training, and support for education and outreach activities.

The current LTER program has significant strengths that will contribute to meeting these new cyberinfrastructure challenges. These strengths include the availability of existing long-term data and network-level products, use of community standards for metadata, policies for sharing data, broad experience in ecoinformatics, a history of informatics research, and an LTER Network Office to serve as the focal point for development efforts. Existing partnerships with the National Center for Ecological Analysis and Synthesis (NCEAS), the San Diego Supercomputer Center (SDSC), and the National Center for Supercomputer Applications (NCSA) are positive collaborative strengths, as are new and growing associations with emerging observatory platforms such as the National Ecological Observatory Network (NEON), the Ocean Observatory Initiative (OOI), and the Water and Environmental Research Systems (WATERS) Network. At the same time, however, a survey of LTER sites also identified impediments. Critical issues that require attention include uneven information management and information technology expertise among network sites; diverse forms of data and methods for collecting and managing data; wide variations in network connectivity (particularly at field sites); and inconsistent access to collaboration technologies.

We propose six strategic CI initiatives to support new and existing science activities in LTER:

1. *Build community-based services and a service-oriented architecture (SOA)* - A scalable, community-based, service-oriented architecture will provide data services to ensure secure and efficient access to data stored in site data repositories, as well as provide computational services for numerically demanding analyses and models and for large-scale multi-site experiments that include sensor networks, satellite sensors, and high performance computing.

2. *Build CI capacity to increase data acquisition, management, and curation at the site level* - Near-term goals for increasing LTER sites' capacity for collecting high-quality data and participating in network-wide experiments, integration, modeling, and synthesis activities will require significant enhancements to staffing and technology.
3. *Build CI capacity to increase data discovery, access, and integration* - Advances in data integration require the development of innovative prototype systems utilizing data warehousing and distributed query systems technologies, linked to research in applying knowledge representation and semantic mediation approaches to harmonize heterogeneous data.
4. *Build CI capacity to increase modeling and analysis activities* - Facilitating and coordinating LTER network-wide analysis and modeling activities aimed at understanding and forecasting changes in regional, continental and global dynamics of social-ecological systems will require significant investment in computing services, software development, and staffing. This effort will require developing scalable computing resources, advanced analytical environments such as scientific workflow systems, and a community-based repository for archiving model code.
5. *Build capacity to increase collaboration* - Collaborative work environments allow scientists residing in different locations to analyze, discuss, annotate, and view data using collaborative analytical tools and video teleconferencing. LTER researchers need access to these tools at both central and remote locations.
6. *Integrate cyberinfrastructure into social-ecological research, education, and training* - Integration of new cyberinfrastructure including advanced tools for analysis and synthesis within the LTER research process will require linking centrally-developed training, education and outreach programs to other training resources that can be remotely accessed by scientists, students, and technicians.

Undertaking these initiatives will require significant new investments in people and technology. These investments are the first step towards achieving a fully integrated research network capable of interdisciplinary, multivariate, and multi-site advances in social-ecological understanding and prediction at spatially and temporally meaningful scales.

## 1.0 Introduction

The CI Strategic Plan is designed to support the research agenda of the LTER Network, and in particular the new agenda for social-ecological research described in the Network's Integrated Research Plan. The new vision for integrated science is inextricably intertwined with developments in cyberinfrastructure (CI).

Cyberinfrastructure is the term coined by an NSF blue-ribbon committee (Atkins et al. 2003) to describe new research environments *"that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services over the Internet. In scientific usage, cyberinfrastructure is a technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge."* Cyberinfrastructure also includes people and organizations that operate and maintain equipment, develop and support software, create standards and best practices, and provide other key services such as security and user support.

Advancing the practice of collaborative science is the major motivation for advancing cyberinfrastructure. In this document we describe a new era of LTER cyberinfrastructure and expound a vision of facilitating and promoting advances in collaborative and synthetic social-ecological science at multiple temporal and spatial scales by maximizing data flows, information synthesis, and knowledge generation. Key cyberinfrastructure will be needed by the Network to achieve its science mission. The required cyberinfrastructure entails building a significant new capacity (Figure 1) within LTER and demands significant new investments in people and technology:

- People - staffing to meet data management and integration needs to match the foreseen increases in data volume and demand for integrated products, to develop applications and services that will accelerate the pace of synthesis, and to develop and conduct education and training to produce a new cadre of IT-adept ecological scientists and cross-trained informatics specialists.
- Technology - for collaboration; communication; data acquisition and generation; data management and curation; data discovery; data integration; knowledge representation; analysis; synthesis; and modeling.

These investments are required to take the next crucial step towards achieving a fully integrated research network capable of advances in social-ecological understanding and prediction at multiple scales. LTER cyberinfrastructure planning and development must be forward looking not only to address existing needs but also to address as yet unanticipated synthetic science over the next ten years. Research that is interdisciplinary, multivariate, and multi-site at these scales will face many challenges and will require significant enhancements to existing cyberinfrastructure. The successful specification and implementation of this cyberinfrastructure will depend on domain scientists and information specialists and will both rely upon and contribute to informatics expertise and CI systems outside of LTER.

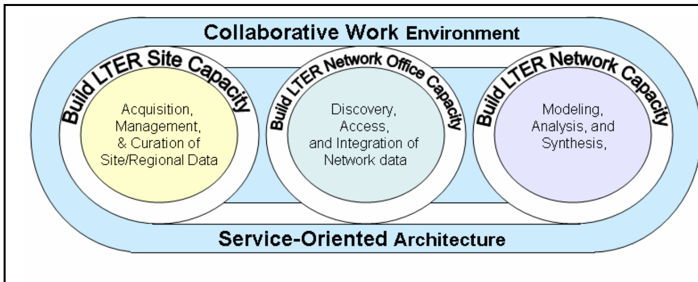


*Figure 1. An integrated vision of a new era of LTER cyberinfrastructure.*

To identify and be better prepared to address these CI challenges, we engaged a diverse group of IT professionals from science and technology centers, large IT development projects, and observatory initiatives to broaden the expertise base for the planning process, facilitate the integration of efforts across programs, and catalyze future partnerships. Interactions focused on four critical function areas that the planners felt would require well-supported cyberinfrastructure:

- Multi-site/Network Experiments,
- Data Integration,
- Modeling, and
- Architecture and Human Resources.

This plan draws from documents produced for each functional area and subsequent interactions with the LTER National Advisory Board and Network Information System Advisory Committee. In the sections below we first (Section 2) present the CI necessary to support a science scenario taken from the LTER science planning effort in order to set the stage for a generalized vision for LTER CI. In Section 3, we present a summary of the current status of CI in the Network based on a detailed inventory and discussion groups. In Section 4 we identify key CI challenges for facilitating Network science. In Section 5, we present a series of 6 strategic initiatives that draw on the strengths of the LTER Network and the IT community to address the challenges identified. Finally in Section 6, we recognize the importance of collaboration with emerging environmental observatories in developing solutions for common challenges.

## 2.0 Cyberinfrastructure Vision and Illustrative Science Scenario

The cyberinfrastructure vision for LTER anticipates integrated activities at the site and network levels to develop critical products that are integrated as key framework components (Figure 2): (1) high-throughput data services that provide high-quality delivery of field-based data products, (2) a service-oriented architecture for a computational environment that allows for the integration of large amounts of multi-site, multidisciplinary data, and (3) collaborative work environments that house comprehensive tools and algorithms for knowledge discovery and data mining, with comprehensive user interfaces that provide tools for easy access, navigation, visualization, and annotation of biological information.

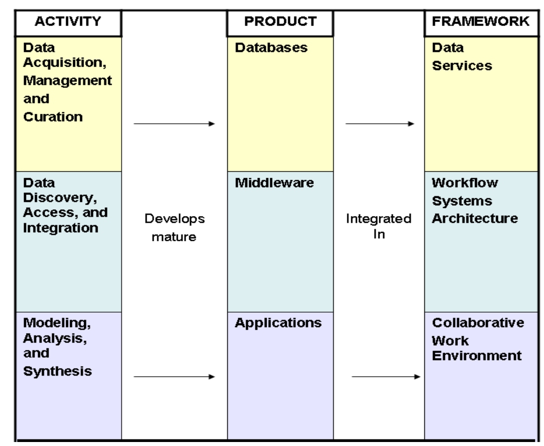

*Figure 2. Diagram of envisioned CI activities, products, and frameworks.*

Three major activities feed into this framework:

1) *Data acquisition, management, and curation*. A framework of data services, tools and expertise that leverages network architecture could potentially support all multi-site studies and experiments. In addition to creating economies of scale, the framework could also provide incentives to researchers to conform to standardized protocols and provide experiment metadata in return for powerful analytical tools and secure data storage. Network personnel could provide design and development support for multi-site experiments such as generating customizable data entry software, designing and curating databases, and creating tools for data quality screening and data query. CI components and personnel at the sites will increase their capacity for collecting high-quality data and for participating in network-wide automated and semi-automated information processing, integration, and synthesis. Information management professionals at the site will participate in both site-specific initiatives and in network, national, and global information systems.

2) *Data discovery, access, and integration*. We envision support for data federation through traditional data warehousing approaches as well as through the implementation of emergent techniques relying on knowledge representation and semantic mediation. A typical network project will involve data sources and users distributed among various institutions. Such

projects require a mature infrastructure that allows seamless integration, analysis, storage, and delivery of information to a distributed community of users.

3) *Modeling, analysis, and synthesis*. We envision the establishment of a modeling and analysis support activity that will 1) promote the synthesis of data across the LTER network, 2) promote the improvement and development of analysis tools and models to answer questions fundamental to the LTER mission, and 3) archive and document data and models and their output for the use by the larger community.

To be carried to fruition, these activities require the development of intermediate products (databases, middleware, and applications) integrated within a framework that provides high throughput data services, a service oriented computational environment, and a collaborative work environment.

The following simple study scenario illustrates these needs:

> Chronic nitrogen deposition removes nitrogen limitation on biotic activity, resulting in diverse unintended consequences to terrestrial, freshwater, and marine ecosystems including changes in plant community composition. To test the hypothesis that changes in N deposition, night-time warming, and precipitation drive long-term changes in plant community composition, and to forecast future continental scale patterns of change, an integrated program of multi-site experimental manipulations and modeling is designed. Changes in N deposition and night-time warming are characterized as "presses." Changes in precipitation are considered "pulses."
>
> To pursue a program of experimentation linked to modeling, background information is needed on precipitation quantities and patterns, N deposition, and temperature regimes to determine the magnitude of treatments for the study. In experimental plots, data to be collected by technicians include plant community composition, N composition of plant species, and soil nutrients. A network of sensors will collect light flux, soil and air temperatures, soil moisture, and relative humidity. New plant community composition data will be integrated with existing long-term community composition data to evaluate how the community change trajectory in the experimental plots compares to the trajectory of the overall site. Thus, the first-level synthesis in this scenario is the integration of new and long-term data from multiple sites.
>
> The ECOTONE simulation model will be used to synthesize process-based understanding gained from experimental manipulations. The model predicts community composition and will simulate the dispersal, establishment, growth, and mortality of individual plants on a small plot (1 to 5 $m^2$) at variable time steps from daily to annually. Soil water content by depth is simulated daily and effects on plant processes are aggregated to one year. Feedbacks among the vegetation, soil water, and soil structure are included in the current formulation of the model. Validation tests of the model at all sites will serve as a test of the overall hypothesis about the role of climate and N deposition in controlling changes in plant community composition.
>
> To extend the generality of study results and forecast the effects of future changes in climate and N-deposition across N. America, a spatially explicit model (10 x 10 km) of plant community composition must be developed. This requires scaling of mechanisms operating at the scale of individual plants to entire communities. Calibration will be based on reconstruction of temporal and spatial patterns observed at individual LTER sites, while validation will be based on reconstructing historic regional patterns of community change. Linkage with global climate models (GCM) and predictions of future rates of N-deposition will drive forecasts of potential future patterns of plant community composition.
>
> Vast acreages of land are exposed to low levels of atmospheric deposition, and hotspots of nitrogen deposition occur downwind of expanding urban centers or large agricultural operations, resulting in

regional effects of impaired visibility and haze in exurban areas and national parks. From a social science standpoint there is the further need for an integrated, multi-site experimental manipulation and modeling effort that might include field testing of potential land management practices to reduce the emission and impact of nitrogen deposition. For example, in what ways can prescribed fire, mechanical thinning, or different harvesting regimes reduce decadal nitrogen accumulation? How are the effects of this accumulation and its potential abatement perceived by the public? How do perceptions translate to behavior? Such an experiment suggests several data needs. For example, a compilation of the geography of existing and potential sources and sinks of nitrogen deposition should include prevailing wind patterns, airshed boundaries, and point sources along with knowledge about the current intensity of land use and historical changes in this intensity since 1960 (e.g., animal processing, transportation corridors). Spatially attributed, high categorical resolution land management, land use history, and attitudinal and perceptual data to differentiate among social groups is needed, as well as parcel data (attributes and geography) and high resolution, multi-spectral imagery for the study area.

*Cyberinfrastructure requirements.* This study scenario begins with synthesis of site-based information that will require more resources at the site level for data acquisition and management, and quickly grows to a scope and complexity that will require integrated CI across all involved sites to be successful. Most obvious of the CI drivers for this scenario are 1) the coordination and collaboration that will be necessary to do the science, and 2) the integration of vast amounts of data derived from historic data collection efforts; from new data collections, including sensor networks; and from multiple distributed sources. In addition, support is needed for multi-site experimental data acquisition including the implementation and management of a complex array of sensors.

The second phase will require the development and parameterization of an integrative model from distributed data sources. The validation of the model will also require multi-site data collection and integration capabilities. The final phase represents new model development to extend the scale of predictive results. New algorithms may be required to extract plant community composition from satellite data. New tools may also be required to validate land use histories and practices and explicit patterns of plant community change. The collection and integration of social science data with ecological data poses additional challenges and relies heavily on spatially explicit GIS and remote sensing data. The multi-site science coordination from design to analysis would be greatly enhanced by the use of collaboration technology.

As it increases in scale, this study requires support for a substantial and integrated cyberinfrastructure. Using the capabilities provided by existing cyberinfrastructure, this study would proceed slowly, requiring months to years to accomplish, if it could be done at all. Enhancements to cyberinfrastructure would greatly facilitate this effort and perhaps even make it possible. A community-based, service-oriented architecture will provide secure and efficient access to site data repositories and satellite data, and provide access to computational services and high performance computing for running models. This architecture would provide a seamless and fault-tolerant environment to allow linkage with global climate models and allow access to predictions of future rates of N-deposition and forecasts of future plant community composition.

## 3.0 Current Status of LTER Cyberinfrastructure

In developing a strategic plan for new LTER CI we took into consideration the existing status and strengths of the Network. The LTER Network is particularly well-suited to take on the challenges presented by the Network's new science agenda because of its many existing CI strengths that are absent or nascent in similar networks. These include:

- Long-term site data that are rich, extensive, well-documented, and online.
- Network-level products that have been developed to facilitate integrative, cross-site research include a network-wide database catalog; network-wide databases for climate and hydrology (ClimDB/HydroDB), for site descriptions (SiteDB), and for bibliographic references; and substantial collections of LTER-wide remotely-sensed imagery.
- Community standards that have been developed and adopted for metadata (Ecological Metadata Language, EML) and site information management; the LTER Network has been the first and largest adopter of metadata standards in the ecological community, and has set standards for site information management systems that have been peer-reviewed and vetted by the ecological community.
- Open data policies have been developed and adopted for release, access, and use of LTER data that clearly define user and provider requirements.
- Wireless sensor networks at a number of sites are providing testbeds for the development and deployment of environmental sensor technologies.
- A Network office is funded and charged with support and leadership in informatics and in computing and communication infrastructure.
- The diversity of knowledge and approaches in the LTER IT community has generated diverse, innovative informatics solutions.
- Informatics research includes scientists committed to network-level ecological research.

## 3.1 Strategic Partnerships

LTER's history of interaction and cooperation has helped to keep LTER IT efforts community-oriented and informed. This strength derives in part from strategic partnerships with national centers and collaborative efforts with ecoinformatics partners:

- The National Center for Ecological Analysis and Synthesis (NCEAS) has been very productive in advancing informatics capabilities for the ecological community and will play a critical role in developing and supporting cyberinfrastructure for synthesis. In particular, NCEAS has been involved in developing tools for generic access to ecological data and will play an increasing role in training and improving technical capabilities of users engaged in synthesis and analysis at NCEAS.
- The San Diego Super Computer Center (SDSC) has established collaborations with the LTER community and provides expertise on information management technologies relevant to observatory networks. SDSC has sponsored training workshops in technical areas of interest such as web services.
- The National Center for Super Computer Applications (NCSA) has been a key collaborator in the development of proposals addressing cyberinfrastructure needs of the LTER Network. In particular NCSA provides critical expertise in Grid architecture and related technologies.
- The LTER Network has strong connections with the emerging National Ecological Observatory Network (NEON), which will provide new partnerships and leveraging opportunities for co-developing and sharing cyberinfrastructure solutions; five LTER sites

are among the 20 preliminary NEON backbone sites, and several others are preliminary gradient sites.

## 3.2 The LTER Network Information System

For the past several years the LTER Network has been developing a Network Information System (NIS) to accelerate the generation and use of data and synthesis products resulting from cross-site research. Modules of the NIS, as it exists currently, include the ClimDB/HydroDB climate and hydrology database, the LTER Personnel and Bibliographic Databases, the site description database SiteDB, and the LTER Data Catalog. The NIS strategic plan provides a number of information management strategies that are aligned with the strategies outlined in this CI strategic plan. The primary focus of the NIS strategic plan is on the use of existing data, improving the quality of data and data discovery through the adoption of the Ecological Metadata Language, increasing the quantity of data available through federated architecture, and facilitating synthesis via applications that use the data and infrastructure. NIS will support standardization in the development and management of information content at the sites through guidance, resources, training, and support. NIS includes the development and deployment of applications that accommodate LTER information content, including an on-line data catalog and applications to exploit these data for discovery of information. NIS will support the creation of Network-based synthetic information products through the use of relational database technology, shared middleware, community-based applications, and scientific collaboration.

The Network Information System Advisory Committee (NISAC) is charged by the Network Executive Board to provide guidance for Network CI priorities and policies, and thus guidance to the LTER Network Office team responsible for the development and deployment of the NIS. NISAC also sets site-level requirements for NIS participation. The involvement of NISAC assures that NIS development is driven by current LTER science issues and needs. NISAC is composed of LTER scientists, information managers, and members of the LTER Network Office, and provides a forum for interaction and communication among these groups. NISAC will also assume a critical advisory role for the implementation of this CI strategic plan.

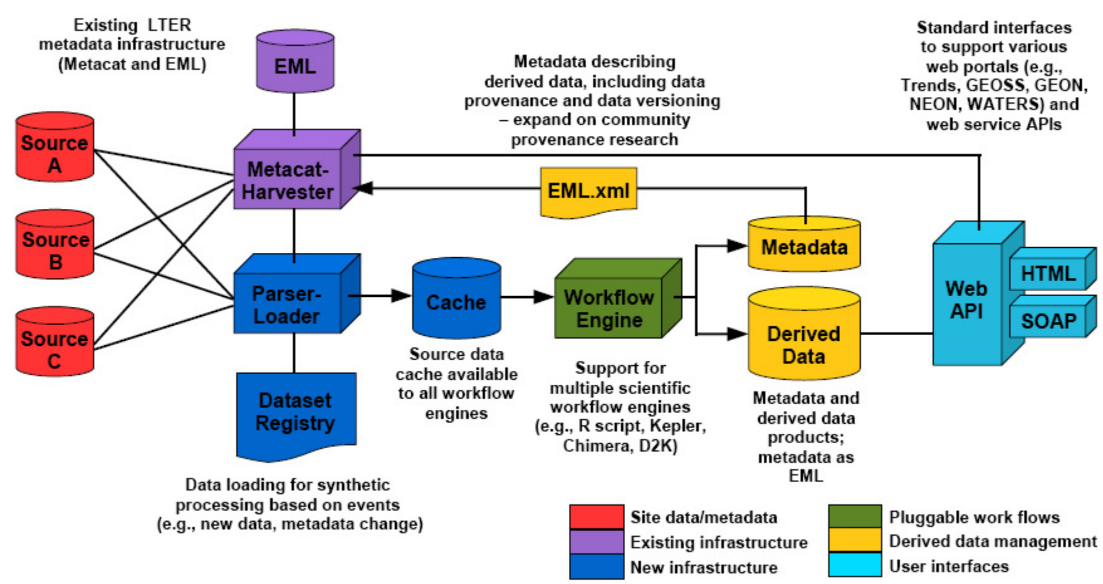


*Figure 3. The Provenance Aware SynThesis Architecture (PASTA) data warehouse framework. The PASTA framework takes advantage of a dynamic archive of Ecological Metadata Language (EML), harvested documentation of LTER site source data, to enable loading and transformation of data sets into synthetic data products. Synthetic products are displayed and accessible via web interfaces and are described in EML and stored in the archive for future use.*

The LTER Network Office team responsible for the design and development of the Network Information System (NIS) has designed and prototyped a data warehouse framework that builds on the successful deployment of Ecological MetadataLanguage, the Metacat repository, and Metacat Harvester. This framework, code-named PASTA for Provenance Aware SynThesis Architecture (Figure 3), is efficient because it builds on existing investments and experiences, integrative because it adopts standard interfaces and approaches, and innovative because it incorporates data provenance and data quality into the design. PASTA is currently being tested as the underlying architecture for the EcoTrends project.

This effort will initially serve the LTER scientific community and collaborators, but it is also seen as a "portal" to the LTER Network for the broader scientific community, students, natural resource managers, policymakers, and the general public. This effort will be continued and strengthened as part of the new LTER science agenda.

### 3.3 LTER CI Survey

As part of the CI planning process we assessed the current status of LTER site cyberinfrastructure. Surveys of LTER sites conducted in June 2005 and February 2007 revealed a very wide range of cyberinfrastructure capabilities among LTER sites (http://lternet.edu/technology). Some of the critical trends are summarized here:

- There is enormous diversity in the available expertise for information management and information technology at the sites, ranging from a quarter to more than 3 FTE supported by LTER funding with as much as an additional 7 FTE from external sources. Most sites receive institutional support for their computational infrastructure, although this varies from email support to more "data center" like operations. Information management tasks range from system administration and user support to software development and web design with the majority of time spent on general site data management activities. Additional information management personnel and training were seen as the most important need for sites to allow them to participate fully in Network-level science.
- LTER site data span a wide variety of forms, from remote sensing data, streaming sensor data, and automated shipboard systems to manually recorded field data. All sites have embraced standards for information management, particularly the implementation of structured metadata in the form of Ecological Metadata Language (EML), although complete metadata documentation and methods for online data access are highly variable across the Network.
- LTER host institutions are generally well-connected to the internet but field sites are highly diverse in their bandwidth and service quality. Fewer than half of LTER field sites have high speed internet connections and wireless infrastructure to support sensor networks.
- Several sites are actively developing and deploying wireless sensor networks to routinely sample and communicate information on measures as diverse as trace gas fluxes, animal movement, and water column chemistry.
- LTER site scientists collaborate on IT issues with domain science centers, and some sites are collaborating internally or externally with computer scientists on IT issues.

- LTER researchers at host institutions generally have access to shared video teleconferencing capabilities but access on individual desktops, in conference rooms and at field sites is sparse: only one third of LTER sites have video conferencing capability of any form at the site. Other collaboration tools are generally not used, and email is almost exclusively the electronic collaboration tool of choice.
- Conventional statistical and analytical software are in use as the norm across the Network, but few sites use advanced remote sensing, visualization, or project management tools.

"Diverse" is the term that best describes the level of functionality across sites. Survey results convincingly demonstrate the need for significantly expanded technology infrastructure and staffing at LTER sites to support a network-wide scientific effort. The LTER Network Office employs several computer scientists and maintains a focus on computing and communication infrastructure, but is not staffed or equipped to address the large project throughput, integration, and data management support required in the coming decade. The LTER network as a whole will require adequate resources and expertise dedicated to cyberinfrastructure to successfully meet a number of critical challenges in transforming the science of ecology to a more highly collaborative and interdisciplinary social-ecological science.

## 4.0 Specific CI Challenges to Facilitating Network Science

There are challenges to exploiting the enormous scientific value of social-ecological data for understanding and predicting the responses of living systems. Collaboration among large groups of distributed scientists is in itself challenging (Hara et al. 2003). Many of the challenges faced by LTER and the social-ecological community are shared by a multitude of other domains attempting to conduct integrative, large scale, computationally intensive science (Maltsev 2006). The challenges in producing high-quality, integrated datasets for synthetic science are immense and long term (Meyer 2006, Stevens 2006). Although the LTER has strengths that provide its scientists an advantage in meeting these challenges, the CI planning process has identified and elaborated specific challenges that include:

1. Acquiring, managing, and curating increasing quantities of data from the network science agenda despite significant diversity in site cyberinfrastructure functionality.
2. Supporting the integration and delivery of increasing quantities of multidisciplinary, multivariate, and multi-site data that will result from new multi-site and interdisciplinary studies, and mediating unavoidable data heterogeneity in site-based ecological studies, including differences in content, format, precision, scale, semantics, and QA/QC. This effort includes explicitly addressing the unique data problems and challenges in using historical social science data and the challenges in using and integrating data from sources outside the Network, such as high-volume geophysical data.
3. Facilitating increasing scientific collaboration organized at multiple geographic scales with dispersed research teams is often not straightforward and requires careful planning to integrate technology (e.g., collaborative work environments, community software tools, conferencing technologies) into scientific practice and to avoid duplication of effort. Integration includes:

a. Facilitation and coordination of CI for LTER network-wide modeling and analysis activities to significantly improve our ability to understand and forecast changes in regional, continental, and global ecosystem dynamics.

b. Developing community-based computing and data services without duplicating efforts elsewhere; and without re-inventing commercial solutions already in the marketplace.

4. Meeting the demand for trained personnel, including cross-trained informatics experts and informatics-adept students and scientists. The high rate of technological change means that training at all levels, from the informatics expert to the individual researcher, will need to be continuously pursued.

5. Ensuring that priorities for CI development and implementation are science-driven, squarely addressing researcher needs. As described in Section 3, the Network Information System Advisory Committee (NISAC) is currently charged with this task and will continue in this role as part of this plan.

## 5.0 Strategic Initiatives to Develop LTER Cyberinfrastructure

We propose to undertake six strategic initiatives in CI to support the major science activities in LTER. These include building network capacity in the critical function areas of data management, data integration, and data analysis, and developing capacity in the crosscutting areas of collaborative work environments, training, and the development of service-oriented architectures. How these strategic initiatives apply to the challenges presented by the LTER science planning effort is readily apparent in some initiatives (e.g., the workforce training initiative), while others leverage the organizational strength of the network to bring a number of applications under a single heading (e.g. data integration capacity).

### Initiative 1: Build community-based services and a service-oriented architecture (SOA)

A service-based architecture serves as the “glue” that holds all the other components together. A scalable community-based service-oriented architecture (Figure 4) can meet the challenges of

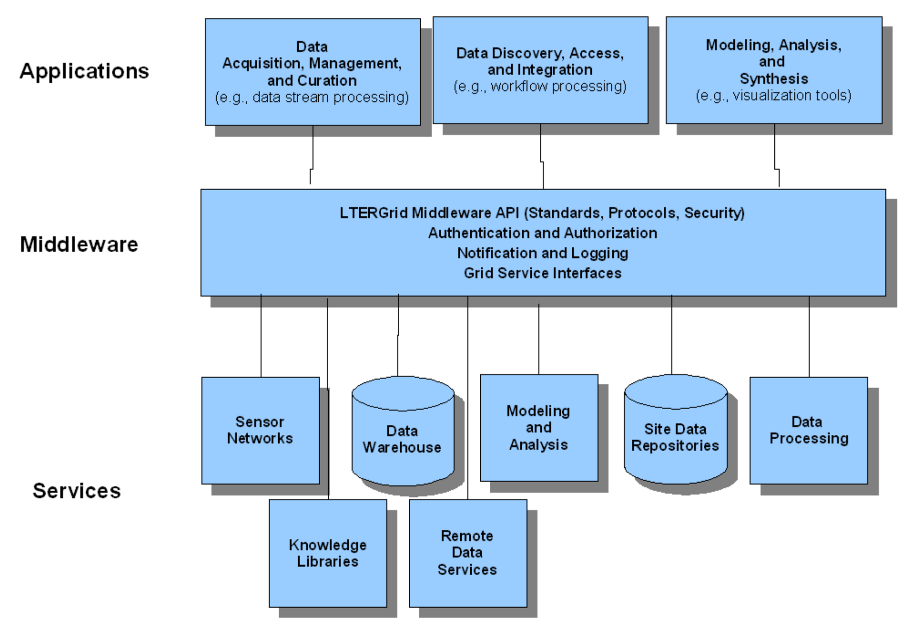


*Figure 4. Schematic of Service-Oriented Architecture (SOA). Communication and security are standardized to present common interfaces to all data for service providers and consumers.*

providing data services that ensure secure and efficient access to data stored in site data repositories, to computational services for numerically demanding analyses and models, and to data from large-scale, multi-site experiments that incorporate sensor networks, satellite sensors, and high performance computing.

Developing and implementing this architecture will require resources and the development of strategic partnerships, including:

- Supporting collaboration with key partners such as NCEAS and NEON to advance the development and deployment of community-based services;
- Supporting integrative software developers and programmers at the LTER Network Office;
- Supporting LTER site participation in the development, deployment, and use of community services.

Achieving a cyberinfrastructure that enables researchers to easily share and exploit current and historical observations depends upon a foundation on which users can discover and access (1) local and remote data, (2) distributed computational resources, including storage and high-performance computational systems, and (3) other collaborators or institutions conducting similar research, all through a secure, fault-tolerant, and seamless process. The framework provided within many of the grid software stacks, including Globus, can be incorporated to implement this vision.

### Rationale

A service-oriented architecture provides a way to expose the functionality of underlying information systems and analytical resources, without needing to implement a centralized system (Alonso et al. 2004). Just as object-oriented programming promotes software reusability by separating essential functionality from the details of implementation, services encapsulate computational and data resources, allowing access through well-structured interfaces that mask the underlying complexities of the supporting systems. There has been an increasing convergence of the standards that support service-oriented architectures and those that support grid computing, and this is leading to new opportunities for the integration of these two approaches.

There are often conflicting requirements between ecological research and grid community developers, with the former accustomed to a "one experiment at a time" approach, and the latter desiring to build systems that handle a large number of tasks simultaneously. Moreover, current grid computing systems are still difficult to learn, not altogether stable, and limited by the technologies available for hardware, software, and programming languages (Hunter and Nielson 2005). It is a challenge to both communities to design better software and use it effectively. With these caveats in mind, one may attempt to select from the available tools and build a robust platform to make routine use of the grid possible. By working with different applications and by addressing common needs and individual requirements, reusable components may be identified without sacrificing the customized environments demanded by users. Collaborative projects with grid application developers are necessary to guide development of this platform and the associated tool development.

### Approach

*Architecture overview*. The grid-based services envisioned for LTER CI will support distributed research sites, sensor arrays, collaborations, and other community services of the Network. To do so will require prototyping community integration through a grid "Point-of-Presence" (PoP) model (Figure 5). Each PoP will provide an interface between the LTER resource and other resources interconnected to the LTER Grid via an Internet2/National Lambda Rail connection. The Site PoP is a combination of networked hardware (server, local disk, and Gigabit network interface) and a software stack consisting of industry standard protocols and applications that provide secure and seamless connectivity from the site to other sites and external resources.

*Software stack*. The PoP software stack (Figure 5) must provide a full complement of services that allow bi-directional connectivity from site resources to any other site or external resources, but at the same time ensure security, fault-tolerance, and an acceptable level of application performance. A crucial service voiced by LTER researchers is security – all access to site data must comply with local authentication and authorization rules, and with the LTER Network Data Access Policy. For this reason, the PoP software services must collect audit information regarding resource usage and the transfer of data.

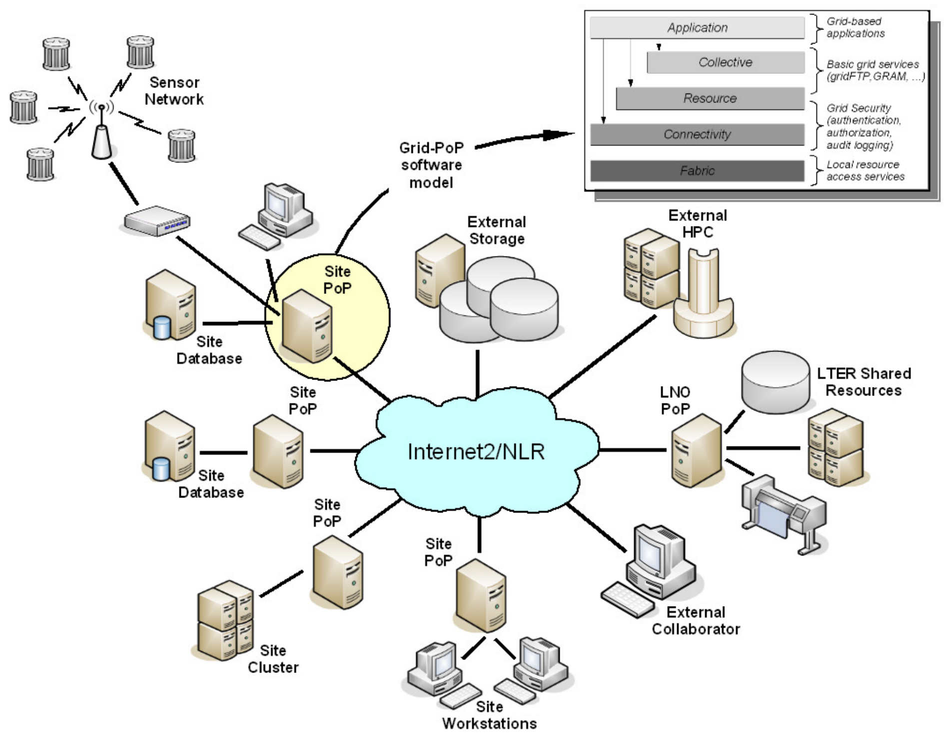


*Figure 5. Schematic of a grid point-of-presence (POP) model architecture for the LTER Network*

## Implementation

First, we will continue developing strategic relationships with supercomputer institutions (e.g., the National Center for Supercomputing Applications and the San Diego Supercomputer Center) in order to leverage their expertise and knowledge in designing and deploying community-based applications and services. As part of this strategic relationship, the LTER Network Office Informatics team will work closely to tailor applications to meet LTER Network and research site needs. We will take advantage of already deployed grid applications that are tested and proven in a production environment as a basis for our design model. Fortunately, many examples are available for scrutiny, such as NEESGrid, BIRN, and TeraGrid. Together with these experienced partners, new partners in NEON and other observatory networks, and a select group of LTER sites that demonstrate their desire to become early adopters, we will prototype a software model to begin connecting and sharing LTER distributed resources.

Second, once the prototype is tested and accepted by the early adopting sites, we will utilize it as a template to deploy to the remaining sites. Following this process will allow us to scale while keeping the site integration time reasonable. In addition, we will strive to bring into the network additional grid-connected resources that are outside the immediate circle of LTER sites such as high-performance computing clusters, visualization tools, and additional off-site storage.

In support of this CI strategic initiative, the LTER Network Office will require additional staff for software development, integration, deployment, and maintenance. It is expected that individual LTER sites will require technical assistance and resources in the initial deployment of necessary software, and also for regular maintenance and system updates as well as some site level staffing resources for enacting the template locally.

**Initiative 2: Build CI capacity to increase data acquisition, management, and curation at the site level**

The challenge of acquiring, managing, and curating increasing quantities of data to support research in coupled human-natural systems requires building capacity at sites by staffing and equipping site information management systems to support data quality, maximum sustainable throughput, and federation of network science data, including data from multi-site experiments. LTER sites need to increase their capacity for collecting high-quality data and for participating in network-wide automated and semi-automated information processing, integration, and synthesis. LTER sites need to assure that information management professionals at the site can materially participate in both site-specific initiatives and in network, national and global information systems. This includes:

- site staffing to support a network information system and maximize throughput of high quality data; this may include network administrators, information managers, programmers, sensor technicians, cross-trained specialists in satellite, sensor, and spatial data, and physical sample archive specialists for maintenance of archive facilities and databases;
- site computing technology to implement persistent data services such as hardware, mass-storage, software, sensors, and physical sample archives;
- LTER Network Office staffing to coordinate development and deployment of standards and web services for site data delivery and site staffing for implementation of services and standards; and
- training Network staff in new technology.

### Rationale

The data collected and managed at LTER sites form the foundation for science at the site, region, multi-site, and Network levels, and hence, meeting each of the articulated CI challenges relies in large part on the capacity at individual sites and coordination among them. New integrative science will demand ready access to online, fully documented data across sites, and LTER site information systems and expertise will likely be leveraged to provide cyberinfrastructure for other research partners within the broader region and in the context of multi-site experiments. Even when data are online and well documented, data integration across sites can be a daunting task. Data management for multi-site experiments has often suffered from a lack of resources and limited and highly variable expertise among experimentalists. Network level experiments will challenge researchers to design and implement functionality both centrally and at the sites to assure data integration.

Increasing volumes of data and new data collection efforts present additional staffing and training issues. Maximizing the throughput of high quality data from field collection to secure storage to centralized access portals is a necessary requirement for supporting synthetic and integrative science. Embedded sensor networks using wireless technologies provide data at new temporal and spatial scales and constitute a new capacity for generating standardized data in multi-site experiments. Maintaining these sensor networks and processing the large volumes of data that can be produced at expanded scales, including automated data screening for quality, will require an increase in staffing at sites. Broader regional representation will require the acquisition of satellite imagery and other spatial data, and coupling human and natural systems will involve collection of new social science data sets such as land use or economics.

The development of properly archived physical samples of specimen collections and reference samples including organisms (plankton, birds, insects, fish, plants, etc.), soils, sediments, and water will demand construction of remotely queriable digital databases of specimen and sample holdings. Voucher collections will allow taxonomic identifications to be verified at a later date and permit documentation of genetic changes in species over time. Reference samples for trace metal, stable isotope, and other analytical methodologies will be similarly important. The advent of new technologies will allow retrospective consideration of samples in order to reconstruct temporal changes in organisms and associated ecosystems, and resources will be required for cross-site comparative analysis of specimen-based materials. Training of site and network staff will be necessary to accommodate new technologies presented by these sensor networks, sample archives, and new data collections.

### Approach

Building this capacity includes obtaining adequate hardware and staffing resources to accommodate the demand for the robust site information systems that are required to collect, manage, and curate data. In order to function as a fully integrated network, the sites must enact network-wide standards and interface with the LTER Network Information System (NIS). This will require sites to develop solutions for exposing site information systems in interoperable ways within the context of a service-oriented architecture. Economies of scale favor the development of tools that can be used across sites to support data acquisition, discovery, access, and integration. Investing in the training of site personnel involved in data collection, information management, and data analysis will provide critical enhancements to the capacity at LTER sites. Funding will need to be available at the sites to support sufficient information management personnel to guarantee that site data meet these requirements so that scientific activities are not constrained by access to data and the metadata necessary to make the data interpretable.

### Implementation

Near-term goals for increasing the sites' capacity to participate fully in the research on coupled human/natural systems will require staffing and technology to:

- enhance data collection methodologies (field data entry systems, sensor network, spatial data capture, sample archive databases, automated QA/QC, automated metadata generation) and train site personnel in the use of new technology;
- improve Ecological Metadata Language documents generated by sites to support interoperability (more content; improved standardization of metadata content across sites, e.g. consistent keywords, units, etc; controlled vocabulary for keywords to permit browser interface to metadata); and
- develop automated access to site data that provides sites with use information (secure web services interface with cross-site authentication; standards for web services content; additional training of site IM personnel to generate site web services).

Mitigating the diversity of current site CI will require developing a critical set of functionality needed at all LTER sites to implement the strategic plan along with a budget estimate for the hardware, software, and staffing needed by the sites to provide this functionality.

**Initiative 3: Build CI capacity to increase data discovery, access, and integration**

To gain ecological knowledge from the anticipated increasing quantity and diversity of data, data must first be curated (evaluated for data quality, linked to ontologies and metadata), integrated (finding and constructing correspondences between elements), and delivered (made available in a form for scientific use).

The challenge of supporting the informed delivery of integrated network data products based on multidisciplinary, multivariate, and multi-site data requires a focused but broad agenda of software development, technical and analytical support, and persistent infrastructure, including:

- LNO staffing to design, prototype, and implement a network information system to integrate site data services; this would include programmers, software developers, and data integration specialists;
- site resources to implement wrappers for site data to conform to specified global schemas necessary for single point of access architecture to LTER site data for specified sets of queries designed by scientists engaged in synthetic multi-site research;
- LNO staffing to provide analytical and technical support for sites in implementing network standards and for the network in utilizing the network information system for synthesis; this may include data and systems analysts;
- funding for collaborative research and working groups focused on mediating data heterogeneity through knowledge representation and ontology development; and
- equipping the LNO to develop and deploy the network information system; which may require adding persistent computation infrastructure in the form of mass storage and computing resources.

### Rationale

Data integration, in the sense in which it is typically used by ecological scientists, is the process of discovering, accessing, interpreting, and integrating data. This process is generally motivated by a scientific question that is not directly addressable by any analyzable data object within our possession. The data integration need arises because we suspect that the data and information necessary to inform our research question exists, and needs only to be assembled into an analyzable object. In the eyes of the ecological scientist, data integration is more a holistic (and currently, largely manual) process involving major unsolved challenges in each area — discovery, access, and interpretation — typically in succession.

Increased capability across the entire scope of the data integration process, including standards for collection, documentation, and communication as well as tools for data discovery and interpretation, involves continued work that addresses the issue at two levels - those associated with data product providers (developing data warehousing, distributed query, and knowledge services for scientific products) and those associated with data consumers (providing easy discovery, access and interpretation of extant data sets).

**Approach**

Data integration herein refers to the entire process of creating a consistent, coherent, and usable scientific resource, a network information system, rather than simply the step of “merging" the data values of independent data sources. Making this process comprehensive and seamless are critical aspects of our data integration agenda.

*Prototyping*. Data integration at the service provider level is defined broadly as combining heterogeneous data sources into a single, unified source and presenting them seamlessly to the user or to applications as a service. Provider-level data integration is complicated by the complexity of the source data and the relationships among them. Data integration at this level is a functional component of either a process known as data warehousing or a part of a distributed query system. Data warehousing involves acquisition, extraction, transformation, and loading of data into an accessible framework. Decisions are made in advance about the relevancy of data sources, the integration approach, attribute mapping, and data quality measures. Although data warehousing is a valuable, presently irreplaceable tool, the process will not be sufficient for the needs of ecology in the future. Distributed query systems differ in that they do not house the data centrally but provide a federated view of the data and always return to the source. Distributed query systems require that mapping decisions be made automatically and therefore rely heavily on knowledge management systems to mediate the queries. We consider prototyping work in both of these aspects of data integration — data warehousing and distributed query systems — to be critical to implementing this strategic plan.

*Research*. Mediating data heterogeneity requires a major investment in applied research and application in knowledge representation and semantic mediation. This includes meetings of strategic focus groups on knowledge management and the addition of software development and data integration expertise to develop ontologies and knowledge services. However, the development of these ontologies still depends on social consensus among scientists — a challenge that involves both social and scientific complexity (Maltsev 2006). The development of new tools and algorithms for mining and clustering existing scientific concepts and terms may provide significant assistance to this process. In exploring possibilities for how heterogeneous data can be discovered, accessed, interpreted, and integrated, the value of developing and promoting standardized approaches to data throughout the LTER cannot be overlooked. Enhanced communication is needed to develop standards so that arbitrary data heterogeneity can be significantly reduced (e.g. in multi-site experimentation, to be sure that there is thorough discussion about methodologies — scale of sampling, methods of treatment application and other aspects of the experiment design). Communication should also lead to common data storage methods, consistent data typing, consistent naming and semantics of variables, minimization of data incompatibility due to spatial and temporal scaling differences, etc.

### Implementation

To gain economies of scale within the new science agenda, the LTER Network Office will play a coordinating role in the integration of research data that requires network-level management by providing data management services that include quality assurance, analysis, and curation. Coordination will involve communication with sites and potentially with data centers for multi-site experiments, and is intended to standardize approaches to data integration by reducing the heterogeneity in data collection methodologies and handling as well as the development of data services within a federated system. The LTER Network Office is particularly well-suited to support this endeavor, but CI investment will have to be substantially increased. In addition, the LTER Network will need to leverage investments in this area that will be made by emerging networks such as NEON.

The implementation of these efforts will be by a small team of developers in collaboration with LTER Information Managers and key informatics partners. This approach has been successfully used in the development of a comprehensive Network-wide metadata catalog and tools and support for implementing the Ecological Metadata Language standard across the Network. This team recently completed a successful pilot project with NCSA to demonstrate the effectiveness of grid technologies on a particular informatics challenge in the network. These efforts will have oversight by a Network level advisory committee (the Network Information System Advisory Committee) consisting of LTER Network Office staff, LTER scientists, and information managers, as described earlier for the Network Information System in Section 3.2.

We will adopt a strategic framework for data integration around the concept of a ‘dataspace’ (see Franklin et al. 2005), an alternative to creating one giant integrated database. The participants in an organization manage a dataspace that encompasses all of the data and information in the organization regardless of its format or location. The dataspace concept structure will allow the LTER organization to model and make available its entire data holdings without exclusion while focusing integration efforts in particular areas of need.

While the LTER Dataspace will provide the conceptual framework, it is necessary to apply a heterogeneous but complementary set of approaches (including application of global schema, automated and manual data warehousing, and knowledge networking) to focus data integration efforts on the highest priority research situations being addressed:

- experimental data where the experiment is designed a priori will benefit from working from a global schema approach;
- post-collection data integration efforts where an ongoing value-added data product is expected should be federated in a data warehouse workflow process, if feasible;
- post-collection data integration efforts where a one-time value-added data product is expected would use manual data warehousing techniques;
- for all data holdings, structural and ontological metadata should continue to be defined and developed to make it possible to do semi-automated data integration for ad hoc analysis; and

- tools for registration and integration of existing databases should be made available, similar to the GEON Portal approach for accessing online resources and the SEEK EcoGrid approach for incorporating semantic mediation and knowledge representation.

**Initiative 4: Build CI capacity to increase modeling and analysis activities**

To facilitate and coordinate network-wide analysis and modeling activities in order to improve our ability to understand and forecast change in social-ecological systems will require significant investment in computing services, software development, and staffing:

- staffing (e.g., programmers, software developers) and increased funding for scientists both at individual sites and at a centralized location that focuses on network-level analysis and modeling activities;
- access to computing services including new hardware technologies, high performance computers, parallel processors, and high storage and high throughput capacity;
- funding for collaboration on software development, including visualization tools, software to link models with different programming languages and the multiple control of linked models, data- and model-based management tools, and network-wide site licenses; and
- equipping the LNO to develop and deploy a persistent archive of data and models which may require adding persistent computation infrastructure in the form of mass storage and computing resources.

This initiative will include support for resources needed by researchers, computational support for analysis and collaborative modeling, and support for an archive for models.

### Rationale

Modeling and advanced data analysis provide critical functions in understanding problems such as ecosystem structure, function and dynamics, responses to climate change, biogeochemical cycling, introduction of exotic species, and changes in human cognition, behaviors, and institutions in response to changes in ecosystem services. Process-based models are an integral part of network-level science activities. Addressing questions that span the variety of ecosystem types across the LTER Network will require models to be integrated with analytical applications, experimental data, observations, remotely sensed images, and spatial databases.

*Advanced Analytical Applications*. The new science agenda will require significant changes in our analytical approaches over the next 10-20 years. These changes will require improved analytical tools, including visualization software, software to take advantage of distributed computing resources, software to link models in different programming languages and the multiple controls of linked models, data- and model-based management tools. The prohibitive costs of "hardening" research software for production-level use makes it necessary to develop robust and scalable tools that can be quickly reconfigured and re-used. Scientific workflows will be needed to document each step in complex analyses so that they can be replicated. Data harmonization and integration will require a host of new analytical tools that will provide semi-automated aggregation and unit conversions, statistical tests for evaluating the effectiveness of integration, and computationally intensive tools that evaluate the impact of decisions in the integration process on the final results of the analysis.

*Modeling*. As the LTER program embarks on questions related to network-level science, models will play an increasingly larger role in the future success of research. Integrated ecosystem, hydrologic, climate, and social science models will be essential for generalizing experimental results and examining interactions between human and natural systems. The nature of new integrated science research questions will likely require the development of a new generation of models to examine non-linear responses, emergent properties, connectivity, and other ecosystem properties, and to couple human and natural systems such as land use prediction models. For example, models can be used to expand the press-pulse dichotomy of drivers to include a continuum of potentially interacting temporal scales. Exploring the importance of spatial variability to ecosystem dynamics across a range of scales, from within-sites to regions and across the continent, is cost-prohibitive through experimentation and best addressed through modeling. Under conditions where transport processes at intermediate scales are important, the extrapolation of fine-scale dynamics to broader spatial extents and longer time periods can only be conducted by spatially-explicit process-based models. Although models are powerful exploratory and predictive tools, the development and use of models for network-level science is currently limited by local technology constraints and resources, the lack of centralized staff and resources dedicated to modeling, and the lack of a formally structured modeling framework.

**Approach**

This initiative will organize and direct computational support of analysis and modeling related activities and identify and collaborate on the development and integration of new analytical tools. A computational framework and collaboration infrastructure are needed to encourage and persistently support modeling and analysis activities. To meet these needs, we foresee a modeling and analysis initiative for the development and implementation of:

*Scalable computing resources*. This initiative will require increased accessibility to new hardware technologies, including high performance computers, parallel processors for some applications, grid technology, and high storage and high throughput capacity.

*Advanced analytical environments*. The use of scientific workflow systems as analytical tools and as a framework around which application and model development and integration can take place is the most promising emerging technology in this area. Scientific workflows are pipelines or networks of analytical steps that may involve, for example, database access and querying steps, data analysis and mining steps, and many other steps including computationally intensive jobs on high performance computing clusters (Ludaescher et al. 2005). The SEEK project is developing an Analysis and Modeling System (AMS) that allows ecologists to design and execute scientific workflows that seamlessly access data sources and services including models, and put them together into reusable workflows. The system is based on Kepler, a scientific workflow system that is community-based and cross-project. This activity will exploit this work and collaborate on specific improvements that will meet the needs of the new LTER science agenda.

Public-private partnerships are also important. For instance, BES researchers are increasingly working to characterize the social and ecological characteristics of all property parcels in the metropolitan region and to link these data to other scales. These analyses are data intensive, often exceeding the computational capacity of existing commercial software. Because many LTER analyses represent the cutting edge of consumer needs, this creates opportunities for commercial vendors to anticipate future market needs and develop products in anticipation of those consumer needs. As a result, commercial partnerships may be valuable for addressing LTER modeling needs.

*Community-based repository*. Archiving environmental data products has become recognized as a vital research practice: it improves our ability to perform new unanticipated analyses and to reproduce results while saving the cost of redundant data collection activities. The same rationale applies to archiving numerical models (Thornton et al., 2005). Archived datasets and models will provide the persistence, provenance, and methodological detail necessary to recreate published results, enabling the synthesis of results across multiple studies and the investigation of new hypotheses. In addition, archived models will allow determination of uncertainties for comparison with results from other models in assessment / policy studies. The model source code will also allow others to see how models treat individual processes.

### Implementation

Development of advanced modeling and analysis capabilities will be a sequential process, focusing first on community-building through a series of workshops. The workshops should draw on both LTER investigators and non-LTER investigators who have relevant synthesis and modeling expertise. The workshop should address questions such as what processes can LTER empirically address across sites now, what modeling capability can be used in the synthesis, and what are the challenges, opportunities, and strategies available for the next 10 years for synthesis and modeling? To sustain synthesis and modeling activities into the future, an ambitious investment is needed to support and train the next generation of modelers. Support for students is needed in the form of targeted graduate student fellowships for synthesis and modeling that includes tuition, competitive stipends, travel allowance for work at multiple sites, and modest research budgets.

Ultimately, we believe that achieving our goals will require the development of a Modeling and Analysis Center that would provide a central location for synthesis and modeling activities and include computational infrastructure, support staff, a director/coordinator, space for visiting scientists (students, staff, researchers), and a venue for synthesis workshops. The center should also be networked to provide year-round accessibility to off-site personnel and include video-conferencing capabilities.

## Initiative 5: Build capacity to increase collaboration

Cyberinfrastructure for collaboration can mitigate distance barriers as research activities increase across multiple scales of geographic distribution and across multiple scientific domains. Efficient, usable, and persistent infrastructure is key to supporting the collaborations and ultimately an integrated research community. By this we mean immediate and continued access to

- staffing for software development and programming of collaborative work environments;
- funding for procurement of video-conferencing technology;
- staffing for software development of integrated analytical tools; and
- funding for procurement of enhanced network infrastructures.

### Rationale

Research activities integrated with CI will enable researchers to work routinely with colleagues at distributed locations (Atkins 2003). To realize this increased capacity demands an understanding of how collaborative work using web-based tools differs from traditional work. It has been shown that geographic distribution can undermine research performance if researchers have not been well-prepared to use collaborative technologies productively. Collaborations that have deliberate social structures, management practices, and frequent contact are more successful (Cummings and Kiesler 2003). Unless the benefits are obvious there will be low tolerance for complicated designs and steep learning curves. The quality of the user-interface, the latency of the network, and the availability of tools are all critical to successful collaboration.

## Approach

Our approach to meeting the increased need for research collaboration will be multi-faceted: procuring and deploying video-conferencing and network technology for immediate use, co-developing and deploying a framework for collaborative work environments, the development and deployment of analytical tools within that framework, and collaboration with socio-technical scientists in order to build effective frameworks and learn from our efforts. Web-based social software holds promise for community-based collaborative frameworks (Figure 6). It is essential that socio-technical expertise in organizational informatics be integrated with this effort to ensure success in meeting this challenge.

*Collaborative work environments*. The development of collaborative work environments will allow scientists residing in different locations to analyze, discuss, annotate, and view data. Access to video-teleconferencing, shared interfaces, community services, and other collaborative tools will allow groups to identify, discuss, and solve scientific problems efficiently. Portal technologies will facilitate communication and common understanding of project tasks and goals through access to data, text, images, etc., among data collectors, managers, analysts, and investigators as well as provide a web content management solution enabling parallel document development (joint authorship). There are, in development, best-practices that take into account the social and technical aspects of collaboration that can help meet these challenge successfully. To develop and implement a CI-based collaborative environment for ecological science requires the integration of ecology with information technology and with expertise in organizational learning. Strategic partnerships and design and programming expertise are required to make this collaborative environment possible.

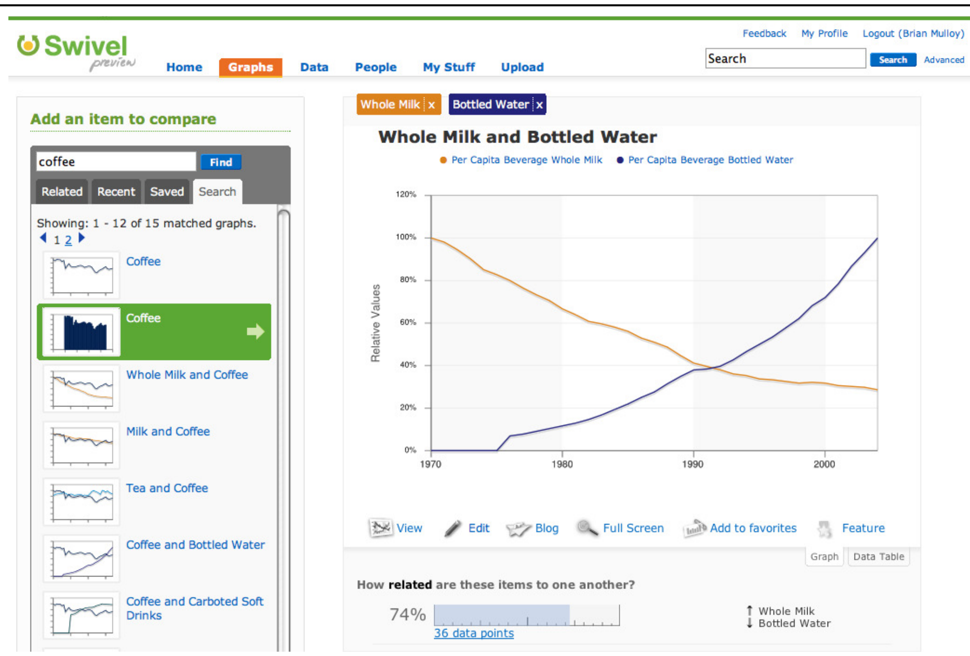


*Figure 6. Web-based social software like Swivel (http://www.swivel.com) empowered with real scientific capabilities could become the workplace for community-based collaborations.*

*Analytical tools*. We must deliver immediate gains through tool deployment while allowing the potential of collaborative work environments to become integrated into normal practice. The researchers must have useful tools, even low-level visualization and analytical tools, to "play with" in order to gain confidence in the use and persistence of the system. Researcher co-design opportunities and programming time are required to further this effort.

*Video teleconferencing*. Common video-teleconferencing (VTC) capability will support multi-site collaboration and information sharing. The LNO has already installed a 48 channel shared VTC facility that provides a basis for support of scheduled and ad hoc meetings from one-on-one communication to large group meetings. Additional hardware and connectivity at the LTER host institutions and sites are needed so that they can use this facility and similar infrastructure to enhance communication across the Network.

*Network connectivity*. Internet2/NLR connectivity at LTER sites will enhance data throughput in the Network and provide site access to Network and other GRID-based resources. To maximize sustained throughput of data and information, a high-level of end-to-end network connectivity from the field sensor network to the investigator desktop, to local and remote data centers will be beneficial. The majority of LTER host institutions, particularly universities, are already linked by Internet2 connections. However, this level of connectivity is not consistent across the Network. For some institutions support is needed to make the link of the last few feet to the local gigaPOP, but for others, collaborative support will be needed to link to a commercial gigaPOP or similar connection in the city center. For the LTER field sites themselves, more than half of the sites have T1 level or slower data throughput. Similar to the needs of the LTER host institutions, additional hardware and network traffic support costs must be met to enhance Network connectivity beyond the current level.

## Initiative 6: Integrate cyberinfrastructure into social-ecological research, education, and training

Integration of new cyberinfrastructure, including advanced tools for analysis and synthesis within the research process, will require training of students and scientists so that their activities will fully reap the benefits of the new technology. There is also a critical training need for technical staff to be kept conversant with new technology and its applications. These challenges can be met by developing a program of workforce training and education with multiple goals:

- provide training in new technologies and methods to information managers and technical professionals engaged in data acquisition and management at LTER sites;
- provide training in the use of advanced informatics tools to natural and social science students and scientists who are engaged in LTER research;
- maintain a cross-trained cadre of information managers who can be quickly deployed with a standard curricula and training materials for working with LTER colleagues and collaborators; and
- develop educational materials tailored to video-teleconferencing, web-based seminars, distance learning, and other paths by which informatics training for educators, students, scientists, and technical professionals can be conducted remotely.

In addition to technological training to support research activities, social-ecological education and outreach will be integrated components of the new network science. Programs will include science education research, engagement of K-16 students in inquiry-based science that integrates social-ecological disciplines and focuses on working with data, opportunities for graduate students to conduct collaborative research within the context of long temporal and broad spatial scales, and efforts to engage the public with broad participation representing our diverse society. In this section we discuss the unique CI needs of training and education and outreach beyond those supporting research activities in general.

### Rationale

Advances in information technologies enable more effective information acquisition, integration, transfer, analysis, and communication, yet the technologies must be harnessed by users who have specific goals in mind and understand which technologies will best accomplish those goals. As the LTER Network is engaged in implementing cutting-edge, enabling CI, the vision of many researchers productively engaging in integrative, interdisciplinary science by easily accessing and analyzing diverse data will only occur with organizational and cultural change that promotes new approaches for designing and conducting science. These new approaches are conceptual (e.g. how do we effectively engage in interdisciplinary research), technological (e.g. what tools can we use to accomplish that research), and social (e.g. how can we convey our findings to other disciplines, policymakers, and the public accurately, appropriately, and with demonstrated relevance).

The new LTER science agenda will produce a tremendous volume of data and information in the not-so-distant future. The societal expectation that we will make good use of that investment argues that a decades-long process of workforce development is not acceptable. Institutional programs designed to train domain scientists in informatics are currently non-existent. These challenges will be addressed by the development and implementation of a training curriculum for graduate students and research scientists. This will result in a generation of students and professional scientists from multiple disciplines engaged in research on complex environmental questions who are able to bring the latest technologies and cyberinfrastructure to bear on the problem of design, conduct, and communication of interdisciplinary research.

Effective use of new technology and the development of innovative, Network-wide CI solutions also demands that training be provided to keep the technical and informatics professionals competent in current and developing technology. This training program is another arena in which partnerships with computer scientists engaged in cutting-edge development can be fostered to facilitate technology transfer. The LTER Network will need to operate at a new level of coordination in order to provide CI for the expanded science agenda, and it is critical that site information managers be involved in the training that such coordination and optimal use of new technology requires. Other technical personnel may benefit from centralized training programs as more LTER sites develop sensor networks and confront the challenges in scaling up the data volumes produced.

Training graduate students, researchers, and technology professionals will support the goals of the network level science program. However, additional training and developments will need to be made to translate technologies and data products into K-16 classrooms and to the public. For example, technology will support the dissemination of diverse (e.g., ecological, sociological, geophysical) data in multiple formats (e.g., real-time, historical, data visualizations). However, achieving educational goals using this new integrated social-ecological research program will require a deeper understanding of what constitutes data literacy within and across disciplines at all levels of education. Similarly we will need to develop effective mechanisms to communicate with and respond to a diverse public.

### Approach

Centrally developed training programs can address the need for cross-trained informatics experts and informatics-adept students and scientists. These programs would include training workshops held at centralized facilities well equipped for hands-on learning as well as other training methods that can be more localized or remotely accessed. To meet some more targeted needs, a cross-trained cadre of information managers will be quickly deployed with training materials. Remote learning environments will be constructed for certain needs that use video-teleconferencing, web-based seminars, and other methods for distance learning. Procedures for evaluation of the training workshops and other materials will be developed. Identification of training needs and development of curricula will involve participation of the targeted user groups.

*Training program for domain-scientists and students.* Training workshops will be provided to graduate students, teaching or research faculty, and research professionals engaged in research on environmental problems. The development of a community of scientists who can use relevant technologies in ecological research can be addressed by 1) providing training in methods and technologies that emphasize information and knowledge management, integration, analysis, synthesis, and dissemination; 2) exposure to example applications where these have been effectively and appropriately applied; and 3) mentoring individuals as they attempt to bring these new approaches into practice. The training program will provide instruction on traditional informatics areas such metadata and database design as well as cutting edge technologies such as embedded sensor networks, scientific workflow software, distributed computing and knowledge representation. The workshops will be structured to provide students with background in fundamental topics before introducing cutting edge technology.

*Training program for technical staff.* Training workshops will be developed for informatics topics identified by both the LTER information management community (for example topics, see Table 1) and IT partners. Each course will include lectures, hands-on labs, and examples of cases where these approaches have been used in environmental studies.

*Special CI developments for education and outreach.* Our CI initiative will address the unique challenges presented in the K-16 educational setting. Approaches such as distance learning technology (e.g., webcasts, linking classrooms, etc.), which are used routinely in undergraduate and research settings must address issues of insufficient infrastructure and technical expertise in the K-12 environment. Some challenges are common to K-12 and undergraduate education. For example, research databases must be tailored to achieve pedagogical goals and must work with educational technology infrastructure. Embedded resources, such as guides to support student inquiry, interactive learning

*Table 1. Potential near-term training needs for technical staff based on a recent survey of LTER sites and Information Managers. Listed tools are examples of relevant tools.*

| Category | Needs |
|---|---|
| Web services | Tools and techniques for participating in Network SOA |
| Spatial data systems | Remote Sensing / GIS / Spatial Data Engines |
| Ecoinformatics | Data management, archiving, and curation; includes EML |
| Data quality | Quality assurance, quality control, quality management |

components, and more engaging graphic interfaces would support the learning community. In short, dissemination of scientific products to educational settings entails dealing creatively with the mismatch between the infrastructures available at K-12 institutions vs. institutions of higher education. In practice, there is a wide continuum of resource availability in schools, and therefore it is necessary to have products and communication available in multiple modes to accommodate this diversity. Finally, research tools (e.g., online or embedded assessment tools) and databases developed for educational research purposes would help to integrate the science education research community.

## 6.0 Collaboration and Integration with other Observatory Networks

Noted throughout this plan are specific instances of potential collaboration with emerging environmental observing networks such as NEON, OOI, and WATERS, and key areas where integration will be crucial in order to leverage synergies and avoid duplication. As a network we are fully committed to participating in the development, testing, and distribution of solutions to the common challenges faced by all environmental observatories.

As detailed in Section 3, many years of CI development both at the site and network levels has provided a rich body of experience and insight with respect to CI deployment. This experience ranges from the development of community-led site-based data standards to network-wide data integration and harvesting technology to the deployment of wireless sensor networks. Wherever possible we will leverage resources, funding opportunities, and education and training activities to further CI needs within the entire family of observatories. At the same time, however, a number of needs will likely remain specific to LTER, at least in the near term, and we will actively pursue meeting these needs in concert with those that are more commonly shared. An example of a more specific LTER issue is the crucial need to integrate and make accessible a rich body of legacy data sets such as those illustrated in the EcoTrends project.

## 7.0 Literature Cited

Alonso, G., F. Casati, H. Kuno, and V. Machiraju. 2004. Web Services: Concepts, Architectures and Applications. Springer-Verlag, New York.

Atkins, D., Kroegemeier, K., Feldman, S., Garcia-Molina, H., Klein, M., et al. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Arlington, VA: National Science Foundation. 2003.

Cummings, J. N., & Kielser, S. 2003. Coordination and success in multidisciplinary scientific collaborations. Paper presented at the International Conference on Information Systems (ICIS), Seattle, WA.

Hara, N., P. Solomon, K. Seung-Lye, D. Sonnewald. 2003. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. Journal of the American Society for Information Science and Technology, Vol. 54(10), pp. 952-965.

Franklin, M, A Halevy, and D. Maier. 2005. From Databases to Dataspaces: A New Abstraction for Information Management. SIGMOD Record, Vol. 34(4), pp. 27-33.

Hunter P., Nielsen P. 2005. A strategy for integrative computational physiology. Physiology (Bethesda) 20: 316-325.

Ludaescher, B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, Y. Zhao. 2005. Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows, 2005. http://users.sdsc.edu/~ludaesch/Paper/kepler-swf.pdf.

Maltsev, N. 2006. Computing and the "age of biology". CT Watch Quarterly. August, 2006.

Meyer, F. 2006. Genome Sequencing vs. Moore's Law: Cyber Challenges for the Next Decade. CT Watch Quarterly. August, 2006.

Servilla, M., J. Brunt, I. San Gil, and D. Costa. (in preparation), PASTA: A Network-level Architecture Design for Automating the Creation of Synthetic Products in the LTER Network. Ecological Informatics.

Stevens, R. 2006. Trends in Cyberinfrastructure for Bioinformatics and Computational Biology. CT Watch Quarterly. August, 2006.

Thornton, P. E. 2005. Biome-BGC: Modeling Effects of Disturbance and Climate (Thornton et al. 2002). Model product. Available on-line [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.