PROSPECTUS

Data from the Long-Term Ecological Research (LTER) Network to Support Synthesis at Site to Continental Scales

Introduction

Long-term ecological data are critical to solving global change questions of national concern and significance. These data provide temporal and spatial context to data collected by existing sites and networks and to emerging national observatories, such as the National Ecological Observatory Network (NEON), WATERS, and the Ocean Observatories Initiative (OOI). As drivers of global change continue to modify our environment, often in unpredictable ways, long-term data will be the only way to distinguish trends in global drivers and ecological responses from short-term variability. However, large amounts of long-term data have been collected that remain inaccessible or difficult to access. Much of these data have been collected at NSF-supported Long Term Ecological Research sites (LTER), but other long-term data sets exist at field stations and marine labs, in individual laboratories, and at Federally-funded research facilities such as USDA Experimental Forests. These data, if made easily accessible, can be used to address critical global change questions.

LTER personnel are uniquely poised to take advantage of the power of long-term datasets, and to link with data to be collected by NEON and other data networks. Information managers at LTER sites have led the development of metadata standards, data dictionaries and software for data integration. LTER sites have made great strides in bringing data sets online over the past decade, but limited resources for data curation and information system development have hindered progress. Consequently, many valuable data sets are stored in non-parsable file formats without comprehensive structural and semantic metadata. Even when data are stored in information systems with detailed metadata, variations in methodology, attribute names, units, measurement scales, code lists and quality control annotation complicate data integration, preventing automated approaches to data synthesis. The LTER Network Office is currently developing a software framework for advanced data integration and synthesis (PASTA - Provenance Aware Synthesis Tracking Architecture) as part of a Network Information System (NIS), which will provide dynamic, centralized access to LTER data for NEON and the broader ecological community. However, this framework will not meet its full potential without a corresponding investment in LTER data curation and information system development at the site level.

Here we propose five alternative models to synthesize primarily LTER data as part of a general approach to synthesis across ecologically-relevant networks, including NEON (Figure 1). All of the models are science-driven, and contain five critical components of synthesis: (1) a broad-based user community for prioritization of tasks and to provide feedback to all steps of the process, (2) centralized efforts to coordinate data synthesis and accessibility across sites and networks, (3) site-based activities for processing source data and creating metadata, (4) activities that promote development and use of derived data products with metadata, and standardized information management products, and (5) tools that allow data and metadata from many sources, sites, and networks to be easily accessed, integrated, and synthesized. Below, we describe each of these needs and provide alternative models/approaches to addressing them.

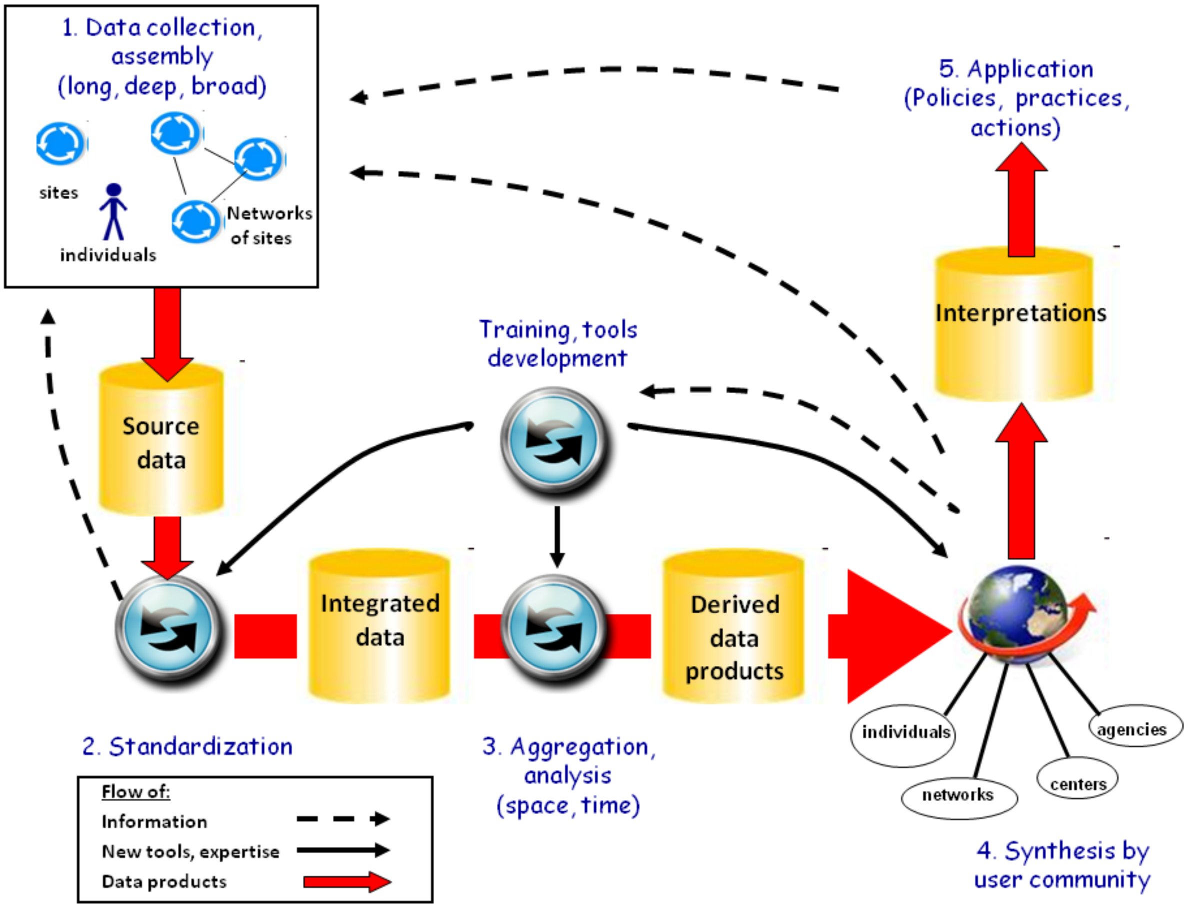

Fig. 1. Our framework for synthesis includes five major activities (shown in blue text) and four classes of products (shown in black text). Modified from: Peters D.P.C. 2011. Accessible ecology: synthesis of the long, deep, and broad. Trends in Ecology and Evolution (in press).

Challenges and Needs for Synthesis

Correct management of LTER data has always been a high priority for sites in the Network. However, because of the historical development of the LTER Network, data management has been largely a site-based activity, which has led to differences in standards and approaches that constitute a barrier to making site data jointly accessible and useful through the Network. The LTER Network has now reached a point at which joint management of site data is seen as the most effective solution to this problem, and the efforts by the LTER Network Office to provide a Network information architecture is one step toward that solution. However, as recognized in the LTER Cyberinfrastructure Strategic Plan, other steps need to be taken to achieve a complete solution. Remedial action needs to be taken to prepare LTER site data and document their formats for ingestion into the NIS framework. This action has not taken place previously because Network data standards have not been available, because the emphasis on site data management has absorbed all available personnel time, and because the lack of a functioning NIS has lowered its priority. Tools to assure the continued processing of new LTER data need to be developed and made available to site Information Managers. Extant data need to be recast to conform to new standards, and data collection systems adjusted to these new standards. Finally, data harvesting into the NIS needs to be implemented and tested at each site. These new tasks require additional short term help to sites both through centralized and distributed sources.

The challenges inherent in synthesizing long-term data to support continental-scale science were highlighted by work in the EcoTrends Project that began in 2006 (http://www.ecotrends.info). EcoTrends was a collaborative effort among state and federal agencies and institutions, led by LTER, to make long-term ecological data easy to access, analyze, and compare within and across sites. Project personnel quickly realized that even within the LTER network, there is significant variation in the quantity and quality of the data and metadata that are on-line, clean, and ready for synthesis and that there is a critical need for coordinated, data cleaning efforts. There are a number of reasons for this variation, including insufficient funding at the site-level to meet all information management priorities. More fundamentally, site-based objectives often take priority over network-level objectives. Our experience suggests that a site-funded model of information management alone is insufficient to facilitate multisite synthesis and that some type of centralized effort to coordinate across sites and with the network as a whole is needed.

Key challenges to synthesis of data across projects, sites, and networks include:

(1) **Prioritization.** A vision for synthesis at site to continental scales is needed to maximize scientific success, push information management forward, and minimize redundancy in data collection and analysis across sites and networks.

(2) **Data heterogeneity.** Network data are in diverse structures that impede understanding and interpretation because they are collected for a specific purpose as part of an experimental or monitoring design rather than for reuse or integration. Format and content standardization are needed to ensure that source data are consistent through time and space and to facilitate integration and synthesis across sites and networks.

(3) **Context.** Interpretation of data often requires additional information to provide context. For example, data on soil texture are likely to be stored separately from biota, or, a time-series of vegetation may not explicitly note the occurrence and description of a disturbance event, such as a hurricane. In addition, directional changes in global-change drivers and/or measurement protocols must be accounted for in interpretation of data from long-term studies. Focused efforts at individual sites within the broader continental-scale vision and following network-level standards are needed to ensure that source data of multiple types are linked appropriately to allow interpretation across sites and networks.

(4) **Accessibility.** A concerted effort by many individuals, state and federal agencies, and other organizations has resulted in the on-line posting of large amounts of source data. However, as noted above, these data often have complex formats and structures at fine temporal and spatial resolutions. A key challenge is to make these large volumes of complicated data accessible and usable by a broad audience through the development of derived data products, tools, and user interfaces

(5) **Flexibility and planning.** Integration systems may take advantage of existing standards (i.e., markup languages) to automate approaches to data synthesis. However, development must be sufficiently flexible and robust to stand the test of time, and incorporate training of future users and developers in these new synthesis approaches.

Large volumes of data have been and continue to be collected that expand the magnitude of these challenges. Such source data are the foundation upon which the rest of the components of synthesis depend. Thus, it is imperative that these challenges be met sooner rather than later as the problems continue to compound through time.

Potential Solutions

LTER experience over the past 30 years suggests that a data synthesis program in support of continental and national-level science questions must be science-driven, and based on coordinated efforts at the site and network levels with active involvement of both information managers and scientists within and beyond the LTER. We propose five major components of a synthesis model, one for each of the major categories of challenges and needs noted above, with a focus on preparing LTER data for use by NEON and other environmental observatories: (1) a science-driven process with priorities derived from a broad user community; (2) a centralized effort to coordinate across sites both within the LTER network and with the broader community to ensure that the highest quality and largest quantity of useful data are made accessible for synthesis; (3) sufficient resources provided at the site level to deal with the large quantities of highly diverse data that have been and continue to be collected; (4) development of network-level standards for derived data products and shared software tools to promote comparisons across diverse sites, variables, and experiments; and (5) training of future developers and users of cross-site, multi-network data for comparisons and analyses.

To implement the components of our synthesis model, we propose five funding options that share certain key characteristics but provide flexible alternatives for achieving our goals. Under each funding option, oversight for the proposed effort would reside with a “Data Council” comprised of LTER Information Management and Domain Science representatives as well as representatives from NEON, USFS, OBFS, and LTREB and possibly other long-term data and synthesis networks. The Council will be responsible for developing dataset priorities (what datasets will be readied in what order) and compliance (what are the standards that make a dataset ready for inclusion). While the full data council will provide oversight, a smaller steering committee (e.g., 3-5 people) would guide and advise the day-to-day operations of this project, which would be headed by a project manager who makes operational decisions. This structure will engage a broad representation of the user and provider communities while providing for effective organization and management of different components of the synthesis model.

The range of funding options to consider includes (from most to least centralized):

- Model 1. The Data Council selects a consulting firm to work with site Information Managers and scientists to discover, modify, and provide data in NIS-appropriate format. Funding is made available to sites to free Information Manager time to work with consultants.

- Model 2. The Data Council selects among proposals from institutional entities (sites; networks; centers; NEON, Inc., the LTER Network Office; etc.) to work with site Information Managers and scientists to discover, modify, and provide data in NIS-appropriate format. Funding is made available to sites to support Information Manager time to work with institutional entities. This model implies that one site (or a synthesis center or research group) could apply for funding to help other sites or other networks provide NIS-appropriate data. In this model, the site or center would be an alternative to the consulting firm in model #1 above.

- Model 3. The Data Council itself oversees a roving Information Management team that visits sites to work with site Information Managers and scientists to discover, modify, and provide data in NIS-appropriate format. The roving team would be drawn from members of the LTER/NEON and broader ecoinformatics community as an alternative to the consulting firm in model #1 above. Funding is made available to sites to support Information Manager time to work with rovers.

- Model 4. The Data Council distributes funds directly to sites based on site-specific needs for providing the required datasets in NIS-appropriate format. For any given type of dataset, funding to a site would depend on the need required to improve accessibility of that dataset: some sites might receive little or no funding while others receive significantly more.

- Model 5. The Data Council distributes funds evenly and directly to LTER sites to discover, modify, and provide data in NIS-appropriate format.

Under any of these options, the following classes of tasks need to be addressed:

Data Council

- Develop a list of national challenges and high level questions
- Recommend priorities among national challenges
- Identify and prioritize necessary kinds of data to address priority national challenges
- Develop sub-teams to test utility of data for synthesis and provide feedback to data and training teams
- Develop partnerships among major data collectors/providers to address national challenges
- Recommend standard formats for data use

Management team

- Manage funds, staffing, subcontracts, to address site data
- Overseee direct assistance to LTER sites for dataset modifications
- Organize tasks to match science priorities
- Provide intensive, on-site technical assistance (e.g., programming, web site structure, database structure) for major data holders
- Coordinate efforts with non-LTER sites and individuals
- Facilitate activities of other teams

Data team

- Identify barriers and solutions to access and standardize data
- Provide specifications for tools to enhance data management
- Check data sets for conformity with standards
- When possible, develop and implement data tools
- Develop and/or adopt data management standards compatible with NIS/NEON

Training team

- Develop short (e.g., 5-day) training courses that will produce standard, documented data and metadata in PASTA-ready format
- Conduct training aimed at LTER, LTREB, OBFS, and USDA Experimental Forest data holders
- From these trainings, provide PASTA-ready data for ingestion into NIS
- Develop web-based training modules for graduate students and postdocs that would be a formal requirement for LTER participation

Analytical team

- Develop specifications for analytical tools to interface with NIS
- Examine existing analytical and data processing tools for suitability (e.g., NEON, DataOne, NBII, commercial suppliers)
- Work with information managers and software engineers to provide scope and cost of additional analytical tools

Why the LTER Program?

Data collection, standardization, curation, and accessibility efforts within the LTER Program began in the early 1980s. This network of 26 sites represents the major ecosystem types of North America, and includes deciduous, coniferous, and boreal forests; arid, semiarid, and mesic grasslands; arctic and alpine tundra, freshwater lakes and streams, coastal and land-coastal interfaces, and urban and agricultural areas. In addition, most LTER sites have expanded regionally to capture a broad range of variability within each ecosystem type and across ecotones. A variety of different kinds of data have been collected from LTER and other long-term study sites through time, ranging from primarily climatic and human demographic data since the 1800s to more recent quantitative assessments of plant, animal, and microbial populations and communities, hydrological and biogeochemical cycles, biodiversity, and disturbance regimes. LTER has the closest linkages with emerging environmental networks, such as NEON, as well as OBFS, LTREB and USDA FS sites; thus the LTER program can lead in the integration of a rich and diverse array of datasets across networks.

LTER personnel are uniquely poised to apply and transfer the interpretive power of long-term datasets to other types of data, including datasets that will come on-line in the future. LTER scientists have expert knowledge of these types of datasets and the questions that can be best addressed by combining long-term data and observations with short-term experimental data and output from simulation models. Information managers at LTER sites have led the development of metadata standards, data dictionaries, and software for environmental data integration. LTER sites have also made great strides in bringing data sets online over the past decade, and have the experience to deal with the challenges associated with the diversity of data that need to be made accessible.

LTER – NEON linkages

Detailed site-based data from experiments and monitoring studies at LTER sites provide a critical foundation for emerging observing systems that are being designed for national-level comparability and, by necessity, must trade off depth and richness of sampling at a site for breadth of sampling across many sites. Complementarity between "drilling down" by LTER sites and standardized national sampling by emerging observation systems will provide ecologists with the data and tools to make significant advances of importance to society. Ready access to LTER data by the scientific community would advance the mission of the NEON program, i.e., to enable understanding and forecasting of the impacts of climate change, land use change, and invasive species on continental scale ecology. Though extensive, the NEON measurement suite is necessarily limited in time, space, and object of study. The LTER Network could provide scientists with essential complementary information (historical and contemporary) to make the best possible use of NEON data.

Though the two programs are different in design and purpose, there are nevertheless important points of contact. For example:

(1) LTER data currently include up to 30 years of intensive measurements (observational and experimental) at individual LTER sites, augmented in many cases by extensive historical, archaeological, dendrochronological and paleoecological records and observations covering the Holocene and in some cases further back in time. These data could provide a critical temporal context for the development of models and the interpretation of trends arising from NEON observations. In particular, land-use and land-cover data from pre-historic to early historic settlement periods, gained with great effort at many LTER sites and difficult to obtain (or unobtainable) from other sources, would directly inform the NEON Land Use Analysis Package.

(2) LTER and NEON sites are partially co-located in the current NEON design. Roughly half of existing LTER sites will be NEON sites; other LTER sites may serve as relocatable sites in the future; and some non-LTER sites that participate in LTER projects such as EcoTrends are also NEON sites (e.g., Santa Rita Experimental Range, Walker Branch Watershed). Where sites coincide, LTER data could directly complement NEON measurements. Where sites do not coincide, LTER data would extend the geographic coverage of measurements in individual NEON domains (each of which has a maximum of three NEON sites). Most LTER sites have a long history of intensive regional studies that could further extend this spatial coverage. Finally, proposed LTER Network activities (such as the future scenarios study) will explicitly include NEON sites and so increase the overlap and complementarity of LTER and NEON data.

(3) Some LTER measurements will align closely with their NEON counterparts, especially in the areas of meteorology, hydrology, and atmospheric exchange to be addressed by the NEON Fundamental Instrument Unit (FIU). But a great strength of the LTER program is the range and diversity of its studies in organismal ecology, biodiversity, phenology, and invasive species. These areas will be addressed by the NEON Fundamental Sentinel Unit (FSU), which is necessarily limited in scope and extent to substrate measurements and indicator species. The

breadth of LTER data would prove critical as scientists seek to confirm trends emerging from NEON observations and explain underlying mechanisms.

The LTER Network Information System (NIS)

The LTER NIS is an infrastructure to enhance the flows of data, synthesis and knowledge about ecological systems to support research collaboration across LTER and partner sites. It consists of information technologies and information products resulting from research activities across sites, and consequently depends on the quality, integrity, and comparability of site data repositories achieved by the implementation of shared standards, software tools, training, and support. The PASTA framework will provide critical cyberinfrastructure components to implement the NIS and is currently under development at the LTER Network Office (LNO).

A fully functional NIS will place considerable demands on personnel resources at individual sites, particularly on site information managers, to assure 1) creation of well documented and quality data products for harvesting into the NIS, and 2) development of standardized approaches and best practices to both improve the quality of LTER data and to facilitate data synthesis and integration. As LTER research transitions from individual, site-based science to broader, more integrative research platforms, so must site information systems evolve to produce the comprehensive structural metadata and quality long-term data now required for participation in a federated database system such as the NIS. Task forces of site personnel will be assembled to develop best practices for information management including quality of LTER metadata documents, data quality assurance techniques, controlled vocabularies for keywords and units, and standardized attributes for common dataset parameters. These efforts will aid preparation of site data for dynamic, metadata-driven loading into the PASTA framework and thereby facilitate the development of derived and other value-added data products. Standardized approaches for such issues as sensor network management and observational data models will be necessary to take full advantage of the PASTA architecture. Successful implementation will be dependent on site expertise and accomplished through both local site efforts and network-wide workshops.

LTER information managers and scientists will also participate in NIS development through an integrated approach with the LNO by 1) informing the development of the PASTA framework components through transfer of expert knowledge and experience to NIS developers, and 2) participating in the development of NIS tools and applications including those that specifically support site information management needs. Information managers will participate in use-case development to determine the functional requirements for PASTA components. The coordination of site and network development efforts will require personnel resources at the sites to ensure that operational steps meet the time-sensitive milestones for NIS implementation. Concurrent development and improvement of site information systems and databases to match the operational development of the NIS at the LTER Network Office is critical to the overall success of the network-wide system.

Prepared by the LTER Network Legacy Data Committee (Peter Groffman, chair; Emery Boose, James Brunt, Ted Gragson, Don Henshaw, David McGuire, Margaret O’Brien, Debra Peters, and Wade Sheldon) on behalf of the LTER Network Executive Board (Phil Robertson, chair).