Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Long term ecological research and information management

William K. Michener [a,*], John Porter [b], Mark Servilla [c], Kristin Vanderbilt [c]

[a] University Libraries, University of New Mexico, USA
[b] Department of Environmental Sciences, University of Virginia, USA
[c] Department of Biology, University of New Mexico, USA

ARTICLE INFO

Article history:
Received 15 November 2010
Accepted 16 November 2010
Available online 22 December 2010

Keywords:
Information management system
Data acquisition
Metadata
Database management system
Data preservation
Data exploration
Data curation

ABSTRACT

The United States Long Term Ecological Research (LTER) Program has supported research in the ecological and environmental sciences for more than three decades. The Program has grown from six to 26 sites and has been the precursor to a worldwide network of International LTER sites. Extracting knowledge from the massive volume of disparate data collected across ecosystems and decades depends upon robust and evolving information management programs at each site as well as a growing, more centralized Network Information System that facilitates inter-site and network-wide data discovery, integration, and synthesis. This paper: (a) reviews the role of policies and governance in the evolution of LTER information management; (b) identifies the components of the human infrastructure that are employed to perform site- and network-level activities; (c) discusses information management functions that are supported at LTER sites grouped by data life cycle components—data acquisition, metadata annotation, incorporation into databases, data exploration/analysis/visualization, and data curation/preservation; and (d) presents the history of the evolution of network-level services within LTER and describes the overall architecture of the Network Information System. Finally, we review the factors that have driven the evolution of information management in LTER over the past three decades and postulate the factors that will guide further evolution of LTER information management during the upcoming decade.

Contents

[*] Corresponding author. University Libraries, University of New Mexico, 1312 Basehart Dr. SE, MSC04 2815, Albuquerque, NM 87131-0001, USA. Tel.: +1 5052772769; fax: +1 5052772541.
E-mail address: wmichene@unm.edu (W.K. Michener).

1574-9541/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2010.11.005

1. Introduction

The United States Long Term Ecological Research (LTER) Program was initiated in 1980 through funding from the National Science Foundation for six initial sites. The program has since expanded to encompass a network of 26 research sites in ecosystems that span latitudinally from the Arctic tundra in Alaska to Antarctic dry valleys and longitudinally from the Moorea coral reef in French Polynesia to tropical rain forests in Puerto Rico (Fig. 1). The network of sites has focused on developing an understanding of ecological patterns and processes at an array of temporal (e.g., diurnal, decadal, century) and spatial (e.g., square meter plots, regional, continental) scales (e.g., Callahan, 1984; Franklin et al., 1990; Magnuson, 1990; Swanson and Sparks, 1990; Kratz et al., 1995; Hobbie et al., 2003).

To date, more than 17,000 peer-reviewed publications have been generated from LTER studies documenting patterns and control of primary productivity; spatial and temporal distributions of populations representing trophic structure; patterns and control of organic matter accumulation in surface layers and sediments; patterns of inorganic inputs and nutrient movement through soils, surface waters, and groundwater; and ecological responses to patterns and frequency of site disturbance (Michener and Waide, 2009). The success of the US LTER Program has led to the adoption of similar approaches in other countries and the establishment of the International LTER Network that includes 38 countries (http://www.ilternet.edu/) that are engaged in long-term, ecosystem-based ecological and socioeconomic research (Gosz et al., 2010).

Research at individual US LTER sites is on average conducted by 18 cooperating investigators and 20 graduate students, as well as by between 10 and 150 other scientists that use research infrastructure available at each site (Gosz et al., 2010). In addition to research, each site has an undergraduate and graduate education program and many offer programs for students and teachers at kindergarten through high school.

Information management is central to the success of LTER. Each site has developed a centralized architecture and staff that support the data life cycle from data and metadata acquisition through incorporation into databases, followed by data exploration, analysis and visualization, and finally ending with curation and preservation. Information management activities at LTER sites have evolved substantially since the 1980's when relatively small amounts of ecological data (e.g., kilobytes) were manually keypunched to today where much larger data volumes (e.g., 10s to 100s of gigabytes) are acquired via both manual and automated approaches (e.g., distributed sensor networks, satellites) on a weekly to annual basis.

The principal objective of this paper is to examine the current state of information management within the US LTER Network. In particular, we first discuss the role of policies and governance in the evolution of LTER information management. Second, we identify the components of the human infrastructure that are employed to perform site- and network-level activities. Third, we discuss many of the information management functions that are supported at LTER sites grouped by data life cycle components—data acquisition, metadata annotation, incorporation into databases, data exploration/analysis/visualization, and data curation/preservation. Next, we present a brief history of the evolution of network-level services within LTER and describe the overall architecture of the Network Information System which has been designed to facilitate inter-site and network-wide data discovery, integration, and synthesis. Finally, we review the factors that have driven the evolution of information management in LTER over the past three decades and postulate the factors that will guide further evolution of LTER information management during the upcoming decade.

2. LTER information policies and network governance

Extracting the maximum scientific value from long-term ecological data requires that data and its supporting metadata be preserved, and that data be available for use by researchers. Both of these requirements seem simple but in practice can be challenging, for both technical and social reasons. Preservation of data requires that metadata (also known as “documentation”, or “data about data”) be complete and accurate, and that data either be actively managed, or kept in good archival formats on persistent storage media (Michener et al., 1997). Making data accessible is primarily a social challenge, but includes technical elements as well.

In the early years of the LTER Network (i.e., 1982–1989), data were managed within individual sites, but generally not shared outside each site. The approach used by LTER researchers, similar to other researchers at the time, was that data use was reserved for the data collectors, or their collaborators. However, over time it became evident that this model was not allowing the most productive use of long-term data. Cross-site comparisons were difficult or impossible because they required contacting each individual researcher involved in data collection (if one could even identify the relevant data collectors). In 1990 two major steps were taken. First, a data catalog containing ten core data sets from each LTER site was published (Michener and Nottrott, 1990). For the first time it was possible to explicitly identify which data were being collected where and by whom. Second, the governing body of the LTER Network laid out a set of guidelines for site data sharing policies, calling on individual LTER sites to compose and publish policies that would address issues of data accessibility (Porter, 2010). Many of the individual site data access policies contained similar elements, and by 1997 the LTER Network had adopted a network-wide policy that built on those common elements. Allowing the data policy to evolve over time was necessary to get the critical buy-in from ecological researchers, and to allow development of an emerging set of ethical principles surrounding data reuse.

The LTER-wide data policy, as modified in 2005, contains sections that define the responsibilities of the data collector, dictating how long data access can be restricted (2 years after collection), and identifying special conditions that may allow additional restrictions for a more extended period (e.g., locations of endangered species, human confidentiality). It also outlines the properties of the required metadata (http://www.lternet.edu/data/netpolicy.html). However, it does not end there. The policy also addresses the responsibilities of users of the data to properly acknowledge the efforts of the data collectors and provides rules regarding the redistribution of data and suggested forms of citation. Proper attribution of ideas and results is central to ethical scientific practice, and the growing acceptance of the need to apply these principles to data has been to a large extent responsible for the willingness of researchers to share data in ways that they would never have tolerated in the past.

The development of metadata (i.e., data about data) standards is an ongoing process involving the LTER Network and its collaborators (e.g., Partnership for Biodiversity Informatics (PBI)). As with data access policies, the development of metadata standards is an ongoing, evolutionary process. It began in 1986 with publication of a workshop volume that included a number of papers that outlined a variety of

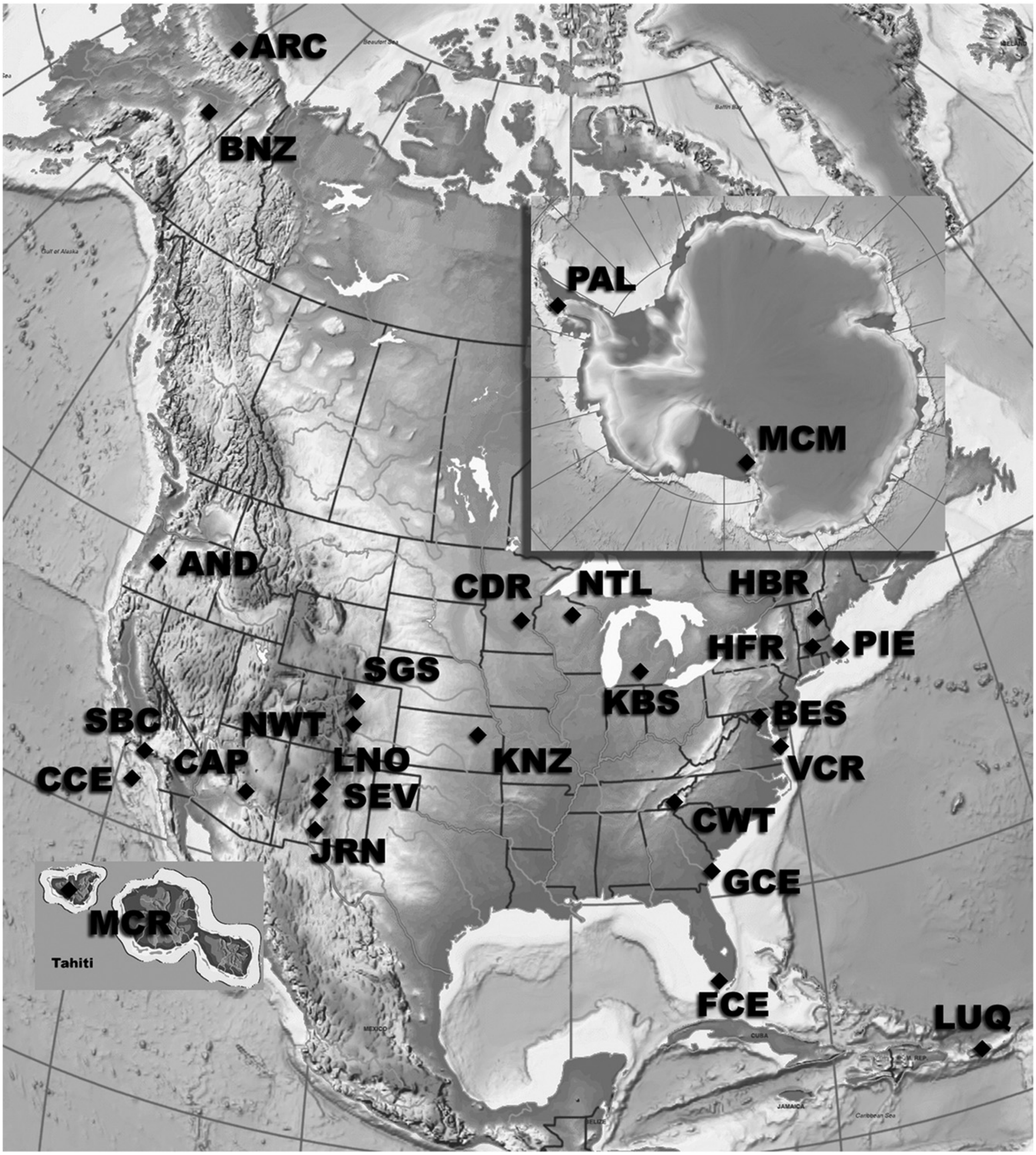


Fig. 1. Location of sites in the US — LTER Network. AND — H.J. Andrews Experimental Forest LTER, Oregon; ARC — Arctic Tundra LTER, Alaska; BES — Baltimore Ecosystem Study LTER, Maryland; CAP — Central Arizona — Phoenix LTER, Arizona; CCE — California Current Ecosystem LTER, California; CDR — Cedar Creek Natural History Area LTER, Minnesota; CWT — Coweeta LTER, North Carolina; FCE — Florida Coastal Everglades LTER, Florida; GCE — Georgia Coastal Ecosystem LTER, Georgia; HBR — Hubbard Brook LTER, New Hampshire; HFR — Harvard Forest LTER, Massachusetts; JRN — Jornada Basin LTER, New Mexico; KBS — Kellogg Biological Station LTER, Michigan; KNZ — Konza Prairie LTER, Kansas; LUQ — Luquillo Experimental Forest LTER, Puerto Rico; MCM — McMurdo Dry Valleys LTER, Antarctica; MCR — Moorea Coral Reef LTER, French Polynesia; NWT — Niwot Ridge LTER, Colorado; NTL — North Temperate Lakes LTER, Wisconsin; PAL — Palmer Station LTER, Antarctica; PIE — Plum Island Ecosystem LTER, Massachusetts; SBC — Santa Barbara Coastal Ecosystem LTER, California; SEV — Sevilleta LTER, New Mexico; SGS — Shortgrass Steppe LTER, Colorado; VCR — Virginia Coast Reserve LTER, Virginia. Map provided courtesy of LTER Network Office, Albuquerque, NM.

approaches to data management (Michener, 1986). In 1992, the LTER Information Management Committee (IMC), based on demonstrations of his "attribute-value syntax" by Thomas Kirchner, began work on developing standard ways of exchanging metadata that would be both human and machine readable. As part of that process, documentation elements used at each of the individual LTER sites were compiled and common elements identified, and in 1994 a "content-standard" for LTER metadata was adopted by the IMC. This nascent standard was then expanded and formalized by the Future of Long-term Ecological Data (FLED) working group of the Ecological Society of America and published in 1997 (Michener et al., 1997). Concurrent with this development were the development of encoding standards, notably eXtensible Markup Language (XML). LTER and its collaborators in the PBI (notably the National Center for Ecological Analysis and Synthesis) then worked on embodying the content standard into a machine and human parsable XML schema as Ecological Metadata Language (EML). EML was adopted in 2002 as the official metadata standard for the exchange of LTER metadata. However, the process did not end there. The information managers within the LTER Network have produced a "best practices" guide for creating EML metadata (http://harvardforest.fas.harvard.edu/data/doc/emlbestpractices_oct2004.pdf), and working groups continue to refine the standards and practices. EML has generally been well accepted and has been adopted by international partners (e.g., Lin et al., 2008).

To support the development of the LTER Network, in particular with respect to information management, the LTER has developed a governance structure (Fig. 2). The LTER Science Council (SC) consists of representatives from each of the LTER sites and is the core of the LTER governance system. The SC has a number of standing committees including the LTER Information Management Committee (IMC). Most day-to-day governance activities of the network are conducted by an Executive Board (EB), which includes an elected chair and rotating membership derived from the SC, along with an elected member from the IMC. Like the Science Council, the IMC has a representative from each of the LTER sites, and elects an executive group (IMExec) to conduct day-to-day governance activities. Not shown in the figure are a large number of ad-hoc information management working groups that draw most of their membership from the IMC, but may include external experts. These groups focus on specific topics (e.g., web services, units registry, controlled vocabularies, web services, quality control and assurance and geographical data) and report to the IMC at an annual meeting. Video-teleconferencing is extensively used by information managers for meetings of all these groups between the, in person, annual meeting.

Two other entities require special attention. The LTER Network Office (LNO) is overseen by the LTER Executive Board. The LNO hosts a variety of network-wide databases (e.g., LTER Data Catalog, All-Site Bibliography, and Personnel). It has a number of ex-officio members of the IMC who participate in meetings and working groups but do not vote. The LNO is also heavily involved in network-wide development efforts aimed at more seamlessly integrating information resources originating at individual LTER sites. To help coordinate the activities of the SC, the IMC and the LNO, the Network Information System Advisory Committee (NISAC) has members from each of those groups. A primary goal of NISAC is to assure that information management activities at all levels facilitate ecological research.

3. Human infrastructure for information management

Different types and numbers of personnel support information management activities at the individual LTER sites and the LTER Network Office, which coordinates network-wide activities.

3.1. Personnel at individual LTER sites

Several roles must be filled at the site level to sustain an information management system (IMS). All LTER sites have at least a half-time information manager (Table 1) who is responsible for overseeing the development of their site's IMS, ensuring compliance with LTER standards, populating databases and web pages, interfacing with scientists to facilitate data and metadata capture, and cooperating with other information managers in developing the Network Information System. The information manager keeps abreast of new technologies and mechanisms for improving discoverability and accessibility of their site's data. Depending on the skill set of the information manager, web programming and custom tool development may be part of their job or may be delegated to student programmers. The information manager has the vision for how all the pieces of the information management system fit together and how the IMS will evolve to better serve site researchers' needs and the needs of the broader ecological community.

The information manager may also be the systems administrator, and responsible for hardware upkeep and supporting local networking needs. In other situations, another individual works part-time to keep servers operational and secure. The need for a systems administrator varies among sites, depending on the degree of cyberinfrastructure support from the host institution.

Management of geospatial data may be the responsibility of the information manager, but about half the LTER sites have at least a part-time dedicated person for geospatial data management. Some sites have a wealth of GIS and remote sensing data while others have little, and this partially dictates whether a geospatial specialist is essential.

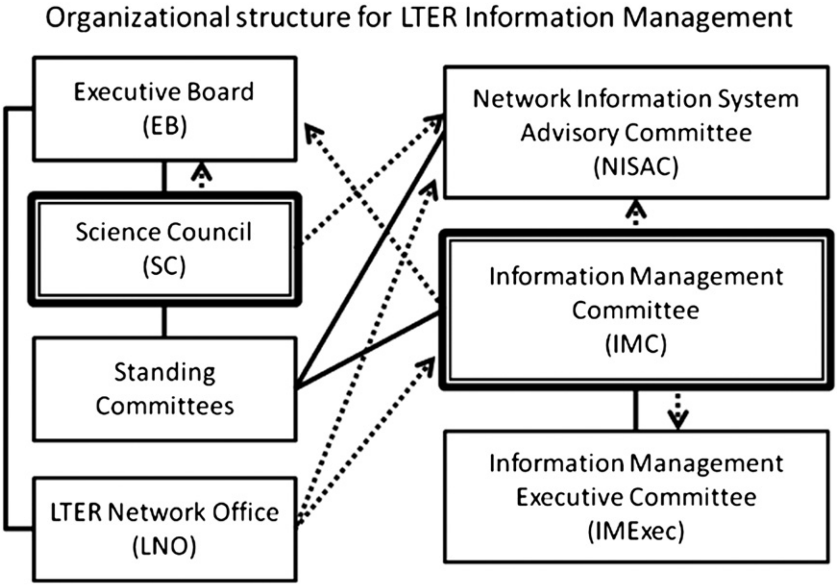

Organizational structure for LTER Information Management
Executive Board (EB)
Science Council (SC)
Standing Committees
LTER Network Office (LNO)
Network Information System Advisory Committee (NISAC)
Information Management Committee (IMC)
Information Management Executive Committee (IMExec)

Fig. 2. LTER governance structure. Boxes with bold boundaries indicate entities where each of the individual LTER sites has a representative. Dotted arrows indicate the source(s) of membership for entities.

Table 1
The roles of IM, Sys Admin, Spatial Data Manager, and Programmer are essential, and may be filled by one or more people. The FTE allocated to each role varies by site, depending on site resources. Sites that are using wireless sensor technologies require additional personnel (positions with *).

Position	FTE	Duties
Information manager	0.5–1.0	• Develops site information management system with databases of site data and metadata • Deploys and updates site webpage • Provides data for cross-site databases
Systems administrator	0.0–0.5	• Operates LAN and file servers • Performs backups • Provides computer user support
Spatial data manager	0.0–0.5	• Manages GIS and remote sensing data
Programmer	0.0–1.0	• Writes software to collect, archive, and process site data
Wireless network and Sensor technician*	0.0–0.5	• Installs and operates the wireless network at the field site • Installs and maintains sensor probes
QA/QC technician*	0.0–0.25	• Monitors visualizations of streaming sensor data to look for problems

While the basic site IMS needs described above can be met by one extraordinary person or, more typically, more than one, there are new research efforts that demand augmentation of site staff. LTER sites are increasingly using sensor networks or flux towers that generate vast amounts of data that need to be accessible to scientists in near-real time. Customized visualization tools are typically needed to support QA/QC of these data. Management and QA/QC of streaming data is a new frontier for many LTER information managers, and additional programming support is often needed. In addition, a wireless network may be employed to transmit the data, and a part-time technician may be needed to maintain this network and associated sensor networks.

3.2. Personnel supporting network activities

It is common at today's LTER sites to find site information managers taking on additional site informatics and data management responsibilities particularly in response to the use of new technologies and an increasing focus on broader-scale synthesis. New streaming sensor arrays and environmental monitoring stations, such as those being deployed as part of the National Ecological Observation Network (Keller et al., 2008), are increasing the rate and volume of locally generated data to be managed by information managers (Ailamaki et al., 2010). When asked to contribute data for cross-site science efforts, and especially where site data are used for Network synthesis studies, site information managers must engage in yet another time consuming effort to prepare data for integration and synthesis. This burden on sites was clearly demonstrated by the EcoTrends project (Laney and Peters, 2006), where much unanticipated human effort was required to cleanse and reformat data into theme-based standards.

To assist in processing and management of data that are collected by LTER sites and to alleviate the burdensome effort required by individual site information managers to ensure that site data are contributed to the Network, the LTER Network Office supports two full-time positions, one to support metadata initiatives at the site-scale and a Network information manager to manage contributed and Network-level data products. The former role is funded through a cooperative agreement with the USGS National Biological Information Infrastructure program, which focuses on expansion and sharing of ecological data through the dissemination of metadata to a network of science-based clearinghouses. Both positions are poised to offset the effort of preparing and submitting data to the Network by the LTER community by helping sites to adopt and understand metadata best practices and standards and by assisting in the transition of data sets from the site into the Network Information System.

4. LTER site information management for the data life cycle

The data life cycle can be described in many ways. Fig. 3 illustrates one view which encompasses two paths: a cycle that starts on the left and encompasses data collection or acquisition, quality assurance/quality control, metadata creation, integration (possibly via a database management system), preservation, analysis and visualization (including data exploration), and publication; and a second cycle that starts on the right and demonstrates re-use of archived data and metadata in conjunction with new integration, analysis and visualization, and

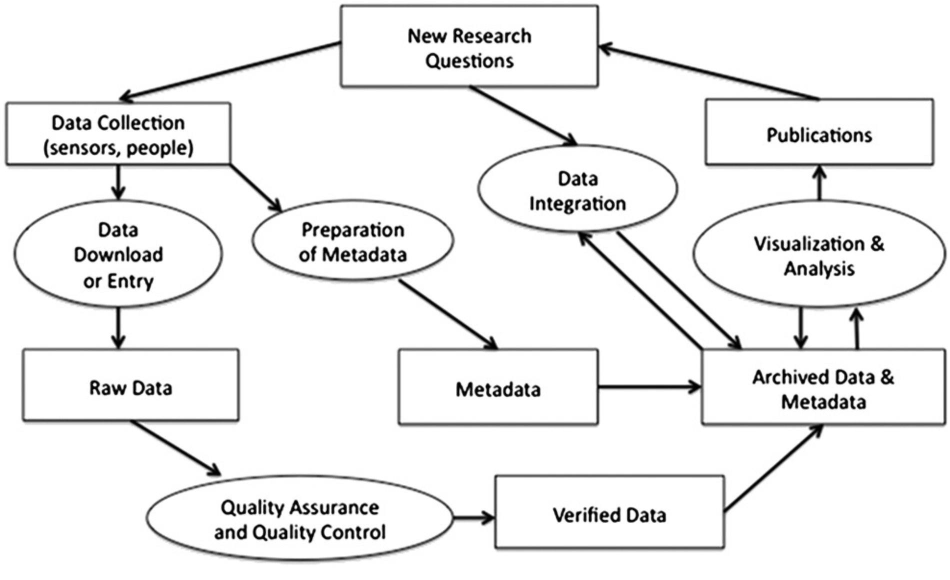


Fig. 3. The data life cycle consists of two major loops. One follows the production of analysis-grade data from scientific question, to measurement, through quality control and assurance, analysis and publication, which then produce new research questions, starting the cycle all over again. However, for data that are archived and shared, a second cycle is possible. In that cycle, new research questions drive the integration of existing archival data, which then allows new analyses that result in additional publications and research questions.

publication efforts. Based on this view, we examine LTER information management activities and group them into several categories: data acquisition; metadata management; database management; exploration, analysis and visualization; and curation and preservation.

4.1. Data acquisition

Many methods for acquiring data are used at LTER sites. Manual techniques are employed for some data collection, but automated data collection is becoming widespread to address questions that require high temporal and spatial resolution. Each method has advantages and disadvantages with respect to data quality.

Many LTER long-term data sets can presently only be collected by hand. Physical sampling such as insect surveys, measuring plants, and monitoring rodent populations require a human to identify species or make measurements. Many quantities measured in labs, such as soil nutrients and precipitation chemistry, may also be recorded manually. Manual techniques for data collection include using data sheets, tape recorders, and handheld computers. Paper data sheets are the time-honored way to record data, but subsequent data entry can result in errors. The big advantage of using paper is that when possible data entry errors are discovered, it is easy to return to the data sheet for clarification. Tape recorders are sometimes used when one of the technician's hands is occupied with, for example, a measuring instrument. A windy day, however, can play havoc with the quality of the recording, sometimes even preventing the data from being recorded. Transcription errors can occur, as well, when the technician listens to the tape via dictaphone and enters the data into the computer. Handheld computers or personal desktop assistants (PDAs) are now commonly used by LTER personnel. This method has the advantage that it avoids transcriptional errors and data can also be downloaded at the end of each day to a network where they can be securely backed up.

LTER scientists have long used meteorological stations or other sensors with associated dataloggers to automatically collect environmental monitoring data (e.g., precipitation, soil temperature), but until recently the numbers of sensors employed was small. Innovations in the types of sensors available to ecologists, coupled with the availability of wireless technology to facilitate data transmission at research sites (Collins et al., 2006) and the decreasing cost of data storage (Sheldon, 2008), have increased the use of sensor networks to measure environmental variables more frequently and over broader spatial scales. Sensor networks have enabled projects such as the Global Lake Ecological Observatory Network (GLEON), an international network of lake ecology observatories that includes lakes at the North Temperate Lakes LTER site in Wisconsin. Instrumented buoys on lakes collect data on key limnological variables and move it in near-real time to web-accessible databases (Kratz et al., 2006) available to all members of the network, resulting in data sets covering a broad spatial domain accessible to a global virtual community. Sensor networks also allow researchers to collect new types of data, such as sounds, or collect data unobtrusively, such as when observing an animal without disturbing it (Porter et al., 2005). Sensor networks have also been employed to measure multiple coupled parameters at high frequency to provide mechanistic process details not available otherwise. At the Sevilleta LTER site in New Mexico, sensors are used to collect carbon flux, water, and temperature measurements simultaneously at high frequency to help scientists understand controls on carbon fluxes (Porter et al., 2009). The quality of sensor data may be adversely affected by electrical noise, mechanical interference, or calibration drift. These factors, coupled with the large volume of data collected by sensor networks, provide new QA/QC challenges.

4.2. Metadata management

A goal of LTER information management is to provide well-documented datasets that can be readily understood and re-used by secondary data consumers. Through the 1980s and until 1997, there were few guidelines related to metadata content and format (Michener et al., 1997). Typically, it was up to the information manager at each site to determine how metadata would be captured, stored, and structured for users. There were as many metadata formats as there were sites, but in 2004 this changed when the LTER information management community adopted Ecological Metadata Language (EML) (Fegraus et al., 2005) as its metadata standard. EML uses eXtensible Markup Language (XML) schema to define content and structure, and consists of several modules to describe different aspects of the metadata (Table 2).

The adoption of EML as the metadata standard for the LTER has led to improved data curation, discovery, and access. The use of XML allows many different types of tools to be used to create and work with EML. EML documents can be generated using general XML editors, as output from databases, or from customized software such as Morpho (Higgins et al., 2002). EML contains fields describing the range and precision of numeric variables and acceptable codes for nominal variables, both of which facilitate automated error checking. EML documents can be automatically harvested from each LTER site and stored in a central database that serves as the back-end for the cross-site LTER Data Portal (http://metacat.lternet.edu/das/lter/index.jsp), from which data from all LTER sites can be searched. Standard XML transformation systems (XSL, XSLT) can be used to transform EML metadata to alternative forms, such as FGDC and ISO-19115, permitting systems using other metadata standards to display LTER metadata in their data catalogs (Fig. 4) and increasing opportunities for discovery of LTER data. Ease of data use is further facilitated by data access information contained in EML that links to data directly from the metadata document.

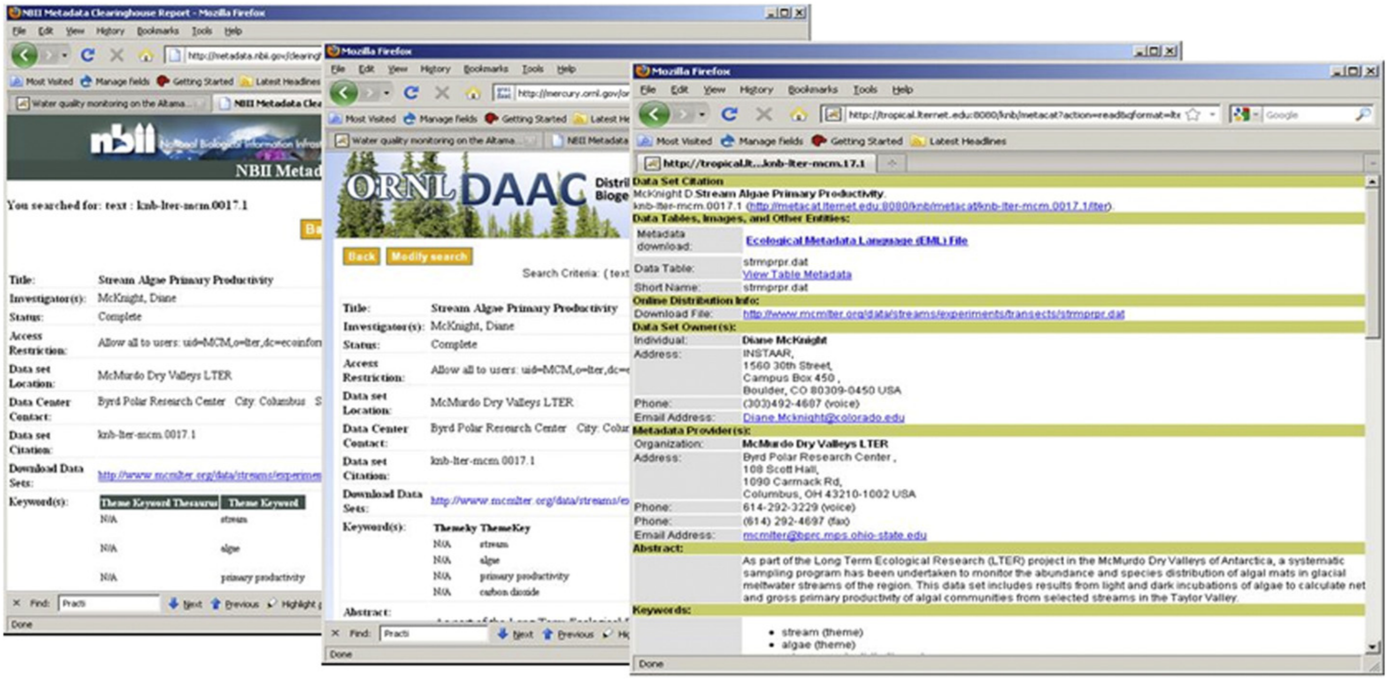
The quality of LTER EML will further be improved by adoption of a controlled vocabulary which is presently under development (Porter, 2010). Use of the controlled vocabulary will improve discoverability of comparable datasets. The LTER Unit Dictionary (Kortz et al., 2009), comprising the set of units in use by LTER sites and the best practices that support them, will also be implemented to ensure standardization of units in data sets. Standardization of units across the LTER will facilitate data synthesis.

4.3. Databases

Databases are one of a variety of tools, along with programming languages and statistical packages, that are widely used in LTER information management. Relational databases that use a Structured Query Language (SQL) interface, such as MySQL, PostgresSQL, Oracle and Microsoft SQLserver and Access, are most commonly used. Built on technology developed primarily for business, not scientific, applications, they provide a rapid and efficient means of organizing, editing, sorting and searching data (Porter, 2000). They are especially useful for the management of metadata, which like much of the data

Table 2
EML modules are designed to describe one logical part of the total metadata that should be included with any ecological dataset.

Module	Content
EML-resource	General discovery level information, such as creator, title, and keywords
EML-party	Detailed contact information for people and organizations
EML-coverage	Geographic, temporal, and taxonomic applicability of the dataset
EML-project	Research context of dataset, such as goals, funding, research question being asked
EML-methods	Field and laboratory procedures used to create the dataset plus quality control procedures
EML-attribute	Name, definition, unit, and domain of variables
EML-access	Level of access allowed to a user or group of users



NBII Metadata Clearinghouse Report - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://metadata.nbii.gov/clearing
Most Visited Manage fields Getting Started Latest Headlines
Water quality monitoring on the Altama...
NBII Metadata Clea
nbii National Biological Information Infrast
NBII Metad
Ba
Find: Practi
Next Previous Highlight
Done

You searched for: text : knb-lter-mcm.0017.1

Title:	Stream Algae Primary Productivity
Investigator(s):	McKnight, Diane
Status:	Complete
Access Restriction:	Allow all to users: uid=MCM,o=lter,dc=ecoinfor
Data set Location:	McMurdo Dry Valleys LTER
Data Center Contact:	Byrd Polar Research Center City: Columbus S
Data set Citation:	knb-lter-mcm.0017.1
Download Data Sets:	http://www.mcmlter.org/data/streams/experimen
Keyword(s):	Theme Keyword Thesaurus / Theme Keyword
	N/A / stream
	N/A / algae
	N/A / primary productivity

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://mercury.ornl.gov/or
Most Visited Manage fields Getting Started Latest He
Water quality monitoring on the Altama...
NBII Metadata
ORNL DAAC
Distri
Bioge
Back
Modify search
Find: Practi
Next Previous Hi
Done

Search Criteria: (text

Title:	Stream Algae Primary Productivity
Investigator(s):	McKnight, Diane
Status:	Complete
Access Restriction:	Allow all to users: uid=MCM,o=lter,dc=
Data set Location:	McMurdo Dry Valleys LTER
Data Center Contact:	Byrd Polar Research Center City: Colur
Data set Citation:	knb-lter-mcm.0017.1
Download Data Sets:	http://www.mcmlter.org/data/streams/ex
Keyword(s):	Themeky ThemeKey
	N/A stream
	N/A algae
	N/A primary productivity
	N/A carbon dioxide
Abstract:	As part of the Long Term Ecological

Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://tropical.lternet.edu:8080/knb/metacat?action=read&qformat=lte
Google
Most Visited Manage fields Getting Started Latest Headlines
http://tropical.lt...knb-lter-mcm.17.1
Done

Data Set Citation

McKnight D.**Stream Algae Primary Productivity**.
knb-lter-mcm.0017.1 (http://metacat.lternet.edu:8080/knb/metacat/knb-lter-mcm.0017.1/lter).

Data Tables, Images, and Other Entities:

Metadata download:	Ecological Metadata Language (EML) File
Data Table:	strmprpr.dat View Table Metadata
Short Name:	strmprpr.dat

Online Distribution Info:

Download File:	http://www.mcmlter.org/data/streams/experiments/transects/strmprpr.dat

Data Set Owner(s):

Individual:	**Diane McKnight**
Address:	INSTAAR, 1560 30th Street, Campus Box 450 , Boulder, CO 80309-0450 USA
Phone:	(303)492-4687 (voice)
Email Address:	Diane.Mcknight@colorado.edu

Metadata Provider(s):

Organization:	**McMurdo Dry Valleys LTER**
Address:	Byrd Polar Research Center , 108 Scott Hall, 1090 Carmack Rd, Columbus, OH 43210-1002 USA
Phone:	614-292-3229 (voice)
Phone:	(614) 292-4697 (fax)
Email Address:	mcmlter@bprc.mps.ohio-state.edu

Abstract:

As part of the Long Term Ecological Research (LTER) project in the McMurdo Dry Valleys of Antarctica, a systematic sampling program has been undertaken to monitor the abundance and species distribution of algal mats in glacial meltwater streams of the region. This data set includes results from light and dark incubations of algae to calculate net and gross primary productivity of algal communities from selected streams in the Taylor Valley.

Keywords:

- stream (theme)
- algae (theme)

Fig. 4. The McMurdo Dry Valleys LTER created a single EML document for the study "Stream Algae Primary Productivity" that has been easily transformed to other metadata standards to make the data set discoverable through several online cross-site data catalogs.

used in a business context (e.g., inventories), needs to be periodically updated in a non-sequential fashion. Ironically, their powerful editing capabilities are less useful for raw data, where appending new data is the most common task, and making changes to existing data is relatively rare. Relational databases apply constraints on the type of content that can be entered (e.g. integer vs. characters) which is a powerful mechanism for assuring data integrity (Vanderbilt and Michener, 2007). Geographical Information Systems incorporating relational databases are ubiquitous for the management of data that have a strong spatial component. Relational databases are widely used at LTER sites to serve behind the scenes to create dynamic World-Wide Web pages over the Internet. Within individual LTER sites relational databases may be linked to web forms provide data searching or reorganization capabilities.

In addition to relational databases, there are specialized databases for the management of XML documents (e.g., eXist) that are used by some sites for the management of EML metadata. However, XML documents can also be generated from metadata stored in relational databases by using programs that query a database, then write out XML documents containing the appropriate tags and structures. In terms of the data cycle (Fig. 3) within LTER, databases often play primary roles in the data entry and metadata preparation steps, but can also be important tools for facilitating data integration and even exploration, analysis and visualization.

4.4. Exploration, analysis and visualization

Visualization and analysis tools help scientists understand and interpret data at multiple levels. At the broader Network scale, exploratory tools help identify relevant and interesting data that may serve as inputs for synthesis or as drivers in numerical models and, often, provide a standardized data product that may be used for cross-site analysis. One early data analysis and visualization application from 1998 is part of the Climate and Hydrologic Database project (Henshaw et al., 2006), "ClimDB/HydroDB", that was developed by the LTER Network, in collaboration with the USDA Forest Service. The ClimDB/HydroDB project federated LTER and USDA Forest Service research sites by developing a data standard that enabled a regular harvest of climate and hydrologic data into a central data warehouse. A simple tool for visualizing the time-series climate and hydrologic data was incorporated in the ClimDB/HydroDB project website through the use of GNUPlot, a burgeoning plotting tool freely available as an open source software application, and provided a means for scientists to easily and quickly evaluate these data from different spatial locations. ClimDB/HydroDB data, and associated metadata, are available for download from the website to those researchers who wish to perform additional further analyses. As a statement of its efficacy, the ClimDB/HydroDB persists today, largely in its original form.

A more recent example of an LTER Network effort that supports online data exploration is the EcoTrends project, which began in 2004, but was not fully realized until 2008 when the EcoTrends Web Portal was launched (Servilla et al., 2008). The EcoTrends Web Portal brings together a broad sampling of standardized long-term time-series ecological data from over 50 research sites, including all 26 LTER sites, into a common data store that provides a rich search interface for discovering data, combined with "on-demand" time-series plots of one or more measurements. Such plots give researchers a graphical view of inherent trends within time-series ecological data and their relationship between multiple variables. Like the ClimDB/HydroDB website, the EcoTrends Web Portal makes all data and metadata available for download by its users. Currently, the EcoTrends Web Portal supports over 15,000 individual data sets.

Innovative software to visualize, analyze and quality control data have been developed by some sites. The Georgia Coastal Ecosystems Data Toolbox for MATLAB (Sheldon, 2002) provides users with tools to visualize legacy and real-time sensor data as frequency histograms, line/scatter plots and map plots and to generate statistical reports. Automated QA/QC is especially useful for screening sensor data, and is accomplished by defining an unlimited number of quality control "rules" for each variable that are automatically evaluated whenever data are imported to generate quality "flags". Interactive QA/QC analysis is supported by providing the user with the ability to manually select, using a mouse, values on a plot to flag. Other LTER sites, such as the North Temperate Lakes in Wisconsin, offer on-demand web plots of sensor data to help researchers see relationships between variables being monitored and also as a QA/QC tool (http://lter.limnology.wisc.edu/sparkling_hourly.html). Other novel visualization approaches include animation of webcam images at the Virginia Coastal Reserve that help scientists better understand processes such as erosion (http://ecocam.evsc.virginia.edu/archive/img1262645903/).

4.5. Curation and preservation

Preserving the utility of data over time is nothing less than a battle against entropy — which dictates that systems tend towards increasing disorder. There are two general approaches to preserving data over the long term. One approach is to store data in relatively simple forms, such as text files, that are based on well established standards and to store data files on persistent media. This approach has the advantage that it requires relatively little expertise or specialized knowledge for a new user to begin to interpret data, a feature that is especially important when there is rapid turnover in personnel. However, the approach provides relatively few services. Tasks such as subsetting data fall to the user. Additionally, rapid technological improvements in storage media make it hard to find a single medium that has a long lifetime. Punch cards, magnetic tapes, 8″ and 5-1/4″ floppy disks were all media of choice in the past. However, it is difficult to find a device capable of reading them now. This leads to the second approach: active management. Here the specific forms of storage of the data are less important than the system of personnel, hardware and software that assures that whatever form the data is in, it will remain accessible. An advantage of this approach is that it works well with complex data structures and allows the use of cutting-edge tools. However, active management can be adversely affected by personnel turnover, and requires a clear migration path for data from the current system into future systems.

Both of these approaches are used at LTER sites, sometimes concurrently, where a database holds a master copy of the data, but it is used to periodically provide updated text files that serve as archival forms. Archival storage of data in proprietary formats should be avoided whenever possible. Spreadsheet and word processing programs, in particular, produce files that often cannot be read by later versions. For example, Microsoft Word 2007 no longer can read document files created by Word versions 1 and 2.

Most LTER sites fall in the middle ground between the “simple form” and “active management” approaches for dealing with metadata. EML metadata documents are text files, so that they are capable of being directly interpreted by a human being, but EML files also contain tags that structure the metadata so that it can also be accessed directly by computer programs without human intervention. EML documents may be maintained using text editors, specialized XML editors, or generated by programs extracting metadata elements from a database and inserting them into an appropriate place in the EML document. Each of these options are used by one or more LTER sites.

Although data sets would appear to be static, through curation, they change and mature over time. Formal quality assurance testing may lead to some corrections, and informal testing as the data is used may lead to others. Additionally, as data matures it may be altered to make it more compatible with other data sources. For example, esoteric codes for species might be replaced with standard codes, or measurement units converted to standard units, or data might be aggregated into different temporal units to aid in the data integration portion of the data cycle (Fig. 3). Good practice dictates that changes should be documented and original, unaltered, copies of data maintained. For example, Borer et al. (2009) advocate making revised copies of data using scripts (programs), so that the raw data remains unchanged, and all changes made in the revised data are documented in the scripts.

5. Network cyberinfrastructure for the data life cycle

The LTER Network cyberinfrastructure for the data life cycle has been evolving for the past 15 years, beginning with technology efforts like the ClimDB/HydroDB project (Henshaw et al., 2006) described earlier, the Data Table of Contents (“DToC”) (Brunt et al., 2002), which amassed LTER metadata records into a web-discoverable database, and the EcoTrends project (Laney and Peters, 2006) and through governance initiatives like the Network Information System Strategic Plan and the Cyberinfrastructure Strategic Plan component of the LTER Decadal Plan (LTER, 2007). The culmination of those efforts has led to the design and planning for the LTER Network Information System (NIS), a centralized metadata and data management system that is being built as part of an LTER community-wide collaboration. Development of the LTER NIS is being led by the LTER Network Office at the University of New Mexico, but includes contributions from sites that range from data products to value-added applications that will analyze and produce knowledge from the data. A primary goal of the LTER NIS is to reduce the human effort required at the site by information managers to identify and harvest site data into this central repository. Unlike the ClimDB/HydroDB model that requires up-front standardization of data and data structures before harvesting, the NIS will rely on descriptive information contained within the Ecological Metadata Language (Nottrott et al., 1999; McCartney and Jones, 2002; Fegraus et al., 2005) document prepared for site data to automatically harvest data in its native format. The EcoTrends Web Portal (Servilla et al., 2008) provided an initial concept design that has guided much of the planning for the NIS.

Funded by the US National Science Foundation, and in part by the American Recovery and Reinvestment Act, design and development of the LTER Network Information System (NIS) began in 2010 and will continue through 2014; a functional prototype that will demonstrate the full data life cycle of LTER data is anticipated for early 2012, with a fully featured NIS expected by 2014. The development progress and information pertaining to the LTER Network Information System can be found at “http://nis.lternet.edu”.

The broad mission of the NIS is to promote advances in collaborative and synthetic ecological science at multiple temporal and spatial scales by providing the information management and technology infrastructure to increase:

- availability and quality of data from LTER sites — by the use and support of standardized approaches to metadata management and access to data;
- timeliness and number of LTER derived data products — by creating a suite of middleware programs and workflows that make it easy to create and maintain integrated data sets derived from LTER data; and
- knowledge generated from the synthesis of LTER data — by creating standardized access and easy to use applications to discover, access, and use LTER data.

The Network Information System will provide the LTER community with a metadata-driven data-flow process for automatically loading data from LTER research sites and making them available through both human and network-based interfaces (Fig. 5). The core services, represented by the Provence Aware Synthesis Tracking Architecture (PASTA) framework (Servilla et al., 2006), is modeled on a Service Oriented Architecture (SOA) approach that organizes autonomous network-aware services into a cohesive, but loosely-coupled system. The PASTA framework is distinguished from the more generalized Network Information System by classifying framework components as critical and enabling cyberinfrastructure that, collectively, provide the services defined by the above mission; the broader NIS includes collaborations from individual sites, as well as “value-adding” applications developed and contributed by third-party

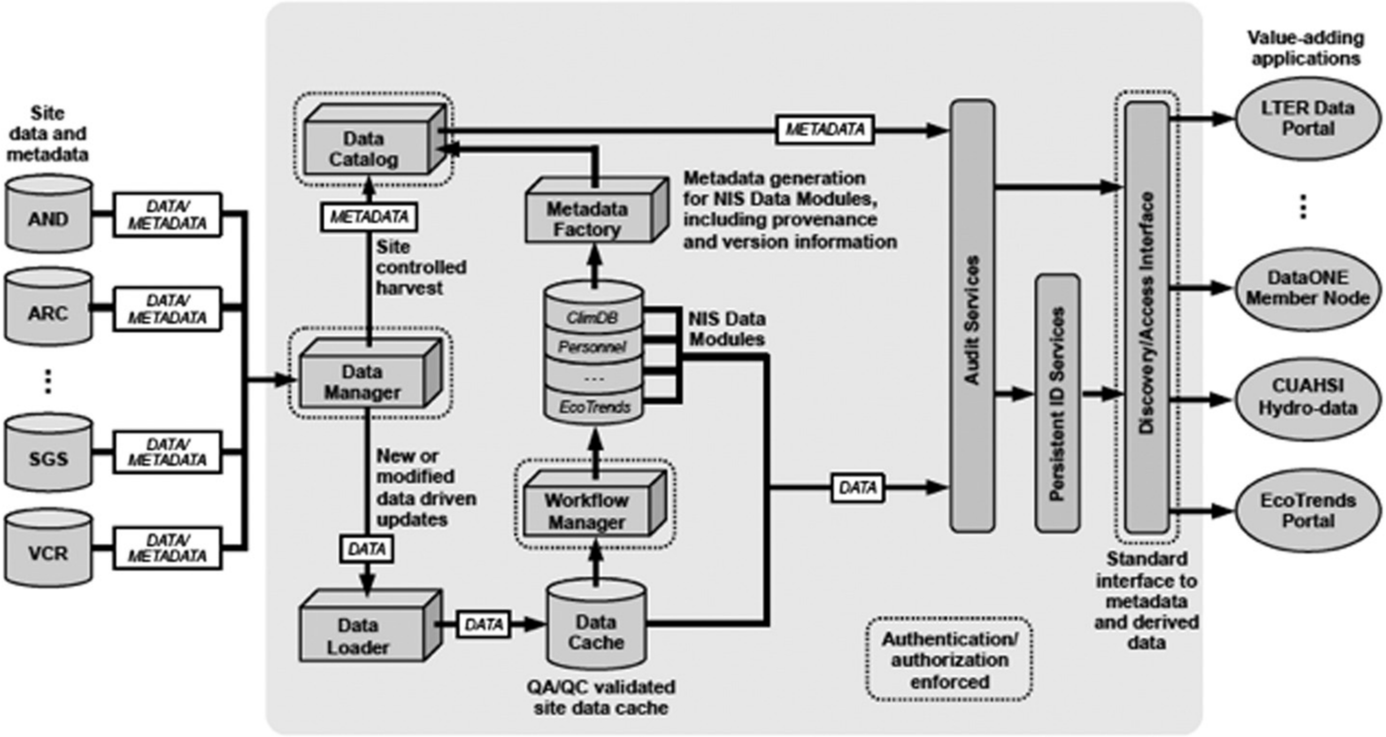


Fig. 5. Metadata and data flow diagram of the LTER Network Information System and PASTA Frame (in gray). Site metadata and data are harvested and made available to workflow execution pipelines to generate derived and/or synthetic data products. A data portal user interface will serve as the primary access point for the general community, while a programmatic interface ("Discovery/Access API") will provide access to more specialized functions of the PASTA Framework.

individuals and groups. All data products within the NIS will have associated metadata, including provenance information where applicable. Initial development will focus on well documented tabular data, which will be followed by more complex data (i.e., spatial and remote sensing data, and/or video data) at a later date. A key goal of the PASTA framework is to simplify the site's burden of participating in the NIS by only requiring that the data be described with complete metadata using the Ecological Metadata Language (EML) standard (Nottrott et al., 1999; McCartney and Jones, 2002; Fegraus et al., 2005) and having the data accessible through one or more Internet protocols. Adoption of any future LTER data standards at the site is not necessary for harvesting into the NIS, but will facilitate future analysis and/or synthesis of such data by applications internal and external to the NIS. As an exemplary value-adding application, site data that is harvested into the NIS will be available to the community through the LTER Data Portal, a web-browser based interface for discovering, accessing, and exploring LTER data products.

The PASTA framework is comprised of nine loosely-coupled services, as illustrated in Fig. 5, that include: 1) Metadata Management Suite (Data Manager and Data Catalog), 2) Data Management Suite (Data Loader and Data Cache), 3) Workflow Management Suite (Metadata Factory and Workflow Manager), 4) Audit Services, 5) Persistent Identifier Services, 6) Identity Management Services, 7) System Monitor, 8) Discovery/Access Application Programming Interface (API), and 9) LTER Data Portal. In the spirit of SOA design, individual and/or collections of PASTA framework services may provide utility to sites by offering a centralized and off-site data archive and distribution point, metadata and data congruency checks for quality assurance and correctness, automated metadata generation and management, data access and use audits, and a source of workflow and data transformation algorithms.

The NIS will use a modified version of the Ameriflux Network classification scheme to identify different classes of LTER data. Data products within the NIS will be classified from 0 to 4 as described below:

- Level-0 — Site located data (may be raw or modified by the site) that are made available for harvest into the NIS.
- Level-1 — Data that are harvested into the NIS as archived replicates of the Level-0 site data; the Level-1 storage structure may differ from that of the Level-0 structure.
- Level-2 — Data are structurally corrected (e.g., realignment or adjustment of columns to match metadata), but retain the same content where possible of the Level-0 data.
- Level-3 — Data that are qualified and processed into one of the NIS Data Modules. Processing may include unit normalization, new label conventions, adjustment to reporting intervals, and reformatting as necessary to be consistent with the target data product.
- Level-4 — Data that is gap-filled and semantically adjusted to meet the needs of particular synthetic data products.

A generalized use-case scenario for the NIS can be visualized by following the metadata and data flow in Fig. 5. It begins with site information managers interacting directly with the Data Manager service interface to configure and schedule metadata harvests into the Data Catalog and to identify "PASTA-ready" data (Level-0) (i.e., data that are made available to PASTA and conform to the necessary metadata standards). Changes to metadata in the Data Catalog that represent new and/or updated site data will trigger the loading of this data into PASTA by the Data Loader. Once in the Data Cache, the data (Level-1 or Level-2) will be available for synthesis projects to produce value-added derived data by incorporating various workflows into the Workflow Manager. The resulting derived data (Level-3 or Level-4) may become part of a recognized NIS Data Module, which is an agreed upon data model resulting from a science-driven goal. Provenance metadata will be captured during processing of all data products and will be included in the final metadata package generated by the Metadata Factory. Discovery and access of both site-based and NIS Data Module data products will occur through the development of "value-adding" applications that are built upon service interfaces defined by the Discovery/Access API; the NIS Data Portal will exemplify such an application. The Audit Services will support and comply with the LTER Data Policy to track LTER data access and usage. A more detailed description of services follow.

5.1. Metadata management suite

The set of services provided by the Metadata Management Suite will enable LTER site information managers to register site-generated metadata for harvesting into the centralized Data Catalog. Site metadata must be valid XML and conform to the Ecological Metadata Language (EML) standard through established LTER best practices. The Metadata Management Suite service interface, represented by the Data Manager in Fig. 5, will support create, read, update, and delete (CRUD) functionality to the Data Catalog for metadata, in addition to services for setting scheduled or ad-hoc harvest events (metadata harvesting in the current LTER information system only supports scheduled harvesting). The service interface will also provide an advanced query syntax to support the discovery of data through attributes and annotations found within metadata. All harvest events will be forwarded to the Audit Services for recording, which will allow site information managers to generate harvest reports. Another important service of the Metadata Management Suite is to notify the Workflow Management Suite that new or updated metadata has been harvested into the Data Catalog and should be evaluated to determine if it describes "PASTA-ready" data; metadata do not need to describe data for them to be harvested into the Data Catalog.

5.2. Data management suite

Represented by the Data Loader and Data Cache in Fig. 5, the Data Management Suite will be responsible for determining whether a data set described by metadata is "PASTA-ready". Through the use of the "EML Data Manager" library, a collection of tools written in Java to interact with EML described data, the Data Loader will parse data-specific metadata fields, generate a corresponding data structure table in a relational database (the Data Cache), and then attempt to load the data into the database table from the site by using standard network protocols (e.g., HTTP, FTP). A successful data load will result in Level-1 or Level-2 data in the Data Cache. Services of the Data Loader will provide "structural" quality control and error detection of Level-0 data, thus enabling corrections and/or flagging of these data prior to being stored; corrected/flagged data will be characterized as Level-2 data. All errors or flagged data that produce Level-2 data will be reported to the Audit Services component of PASTA; sites will have access to this information through the Audit Services. The Data Loader service interface will also provide metadata/data quality checking capabilities (i.e., ensuring data formats comply with descriptions in metadata) to site information managers for pre-harvest evaluation of their metadata and data.

5.3. Workflow management suite

The Workflow Management Suite consists of the Workflow Manager and the Metadata Factory (Fig. 5), two tightly-coupled services that operate during any metadata/data operation within the PASTA framework. The Workflow Manager provides a repository and service interface to store and execute workflows within the scope of the PASTA framework. A workflow is simply a sequence of computational tasks that perform some action in the PASTA framework. In general, workflows provide a way to document standard processing steps used to generate derived data products and the ability to modify and retrace those steps, but also provide other services such as invoking the Data Management Suite to evaluate metadata that may result in a new data load and/or to initiate the generation of a new provenance-enhanced metadata record by the Metadata Factory. For creating derived data, the level of automation possible and the effort required to develop a workflow is highly dependent on the complexity of both source data and the derived data model. Early efforts to generate workflows will be restricted to simple scripts and pre-written executable programs that reference known data from within the Data Cache.

Most workflow events will result in the Metadata Factory generating metadata using the Ecological Metadata Language (EML) standard format, including provenance information, for all Level-1 and Level-2 data found in the Data Cache and, more importantly, all derived data products (Level-3) that are produced by the Workflow Manager. Each EML document produced by the Metadata Factory will be harvested into the Data Catalog. Provenance metadata for each derived data product will include one or more references to metadata documents (also as EML found in the Data Catalog) that describe the original site-based data set(s) (Level-1 or Level-2) used to generate Level-3 derived data products. A natural language description of the program used to generate the derived product will be included in the metadata, as well as the source code of any script or executable code used to generate the derived data, including configuration and steps that are part of an external workflow package. The Metadata Factory will generate metadata sufficient to recreate a derived data product based on its Level-1 or Level-2 inputs and the computational tasks used in the process. The metadata structure that will be used to store provenance information is the "methods" sub-tree of EML. This sub-tree will contain core elements of the original EML that describes the Level-0 data and Level-1 and/or the Level-2 data that are used to produce any derived data product, including a direct reference to these documents within the Data Catalog — we refer to this reference as a "metadata chain". The "methods" sub-tree will also contain available source code and/or a description of the executable used in the workflow sequence. As with the other PASTA framework services, all Workflow Management Suite events will be recorded to the Audit Service for subsequent reporting.

5.4. Audit services

Audit Services will provide a general framework for recording information related to events that occur within the PASTA framework. Each service that generates an event record may pass this record to the Audit Service, which will then store and archive the record for future reports. Audit Services will support a query syntax for discovering specific records, which will be important for generating reports related to critical events (e.g., data access events for a specific LTER site). Key attributes and events that will be recorded by the Audit Services include user information that is required for compliance with the LTER Data Access Policy, metadata access events, workflow execution events, and user authentication events, to name but a few. An ISO 8601 formatted date/time-stamp will accompany all recorded events.

5.5. Persistent identifier services

A problem that often occurs when digital objects are identified with absolute locations, such as a web URL to a data set, is that the direct link to the object "breaks" when the object is moved to a new location or the domain name of the web server is changed. To overcome the limitation of absolute identifiers, persistent identifiers use a relative identification scheme that maps the identifier to the current location of the object. As such, the user of a persistent identifier can be assured that the identifier will always resolve to the object, regardless of its physical location. The Persistent Identifier

Services will provide persistent identification of (and resolvability to) digital objects (including metadata documents and data sets). In addition to object persistence, the Persistent Identifier Services will ensure that once an object is identified by a unique identifier, no other object within the operational domain may be identified with the same identifier. The use of persistent identifiers will be optional for LTER Network data (Level-0) that resides in a site-based information system and is referenced through a site-generated EML document found in the Data Catalog. All metadata and data found in the PASTA framework, however, will be identified with a persistent identifier.

5.6. Identity management services

Identity Management Services will provide support for single sign-on authentication to access PASTA framework services. Single sign-on implies that a user may authenticate at one, of perhaps many, authentication interface points associated with the LTER Network (including the broader community of service providers) and would not have to re-issue the same authentication credentials at each service that requires user authentication; the credential or token representing the user provides short-lived authentication on behalf of the user to services requiring authentication. Because of the complexity and widespread need of such services, the design and implementation of the Identity Management Services will be developed collaboratively through multiple working groups within the community, including the CILogon project at the University of Illinois Urbana Champaign (http://www.cilogon.org) and DataONE (http://dataone.org) project at the University of New Mexico. All authentication events will be forwarded to the Audit Services for recording.

5.7. System monitor

The System Monitor will monitor and provide system state-of-health information for all PASTA framework services. The System Monitor will consist of a number of component tests that evaluate the state of specific services and/or collections of services. Each test will be measured from a known performance base-line, thus highlighting changes in performance (either positive or negative) of the service(s). The System Monitor will operate as a background process from which state-of-health alerts and reports may be continually generated. These reports will be separate from those available through the Audit Services.

5.8. Discovery/access application programming interface

The Discovery/Access Application Programming Interface (API) will be the standard interface by which external "value-adding" applications will interact with the LTER Network Information System and, more specifically, PASTA framework services. The Discovery/Access API will provide a web-services interface that follows a Service Oriented Architecture (SOA) design, thereby supporting a neutral programming layer that may be used by multiple applications regardless of their hardware and software preference. The Discovery/Access API will support a core set of services to provide discovery and accessibility to LTER data products (Levels 1–4). In addition to the core services, the API will also support user authentication for single sign-on through the Identity Management Services, system information (including, state of health), site-developed applications and tools (e.g., Controlled Vocabulary and Unit Registry), and project specific data services (e.g., CUAHSI Hydrologic Information System (HIS), DataONE Member Node, National Ecological Observatory Network (NEON) functions). The Discovery/Access API will be an abstraction layer that aggregates the specification of all publicly facing service interfaces, while eliminating (or minimizing, at the very least) tightly-coupled dependencies between underlying PASTA framework components from external applications; as such, framework components may evolve as necessary without adversely affecting applications that are built on top of the Discovery/Access API.

5.9. LTER data portal

The LTER Data Portal will be the community's web presence for discovering, accessing, and exploring LTER data products and will serve as an exemplar of a "value-adding" application that utilizes services provided by the PASTA framework. The Data Portal will allow a user to customize their interface by setting preferences for the "look and feel" during their personal web-browser session, including criteria for data discovery and access. The Data Portal will set the standard for key "portlets" that may be reused by other web-based portals (in support of the LTER community "skin" adoption of the current LTER Data Catalog web presence).

6. Conclusion and future directions

Information management within the U.S. LTER network has evolved significantly since 1980 in response to needs of the scientific community as well as advances in information technologies. Some of the key drivers of change in LTER information management over the past three decades have included:

- expansion of the research network to include new sites in different types of ecosystems;
- broadening of the science to incorporating other domains such as the social, behavioral and economic sciences;
- increased scope of scientific enquiry and synthesis that goes beyond single and multi-disciplinary studies engaging a single or small number of investigators to interdisciplinary and transdisciplinary studies that are engaging large numbers of scientists across multiple domains and from throughout the network of sites;
- development and adoption of sensor technologies that expand the scale and capacity of data acquisition capabilities by orders of magnitude such as remote sensing, video, embedded environmental sensor networks, LIDAR, and eddy flux measurement systems;
- rapid and significant improvements in data and information management technologies such as new relational and object-oriented database management systems, as well as the dramatic growth in capability of geographic (GIS) and laboratory information management systems (LIMS); and
- availability of new and improved hardware and software systems that facilitate data exploration, analysis, and visualization, including high performance computing, scientific/visualization workflow systems, and high-resolution video technologies.

We anticipate that many of the same science needs and technological changes will continue to drive advancements in the capabilities and pace of information management. In particular, increased recognition of the connectedness of ecosystems throughout the biosphere and the dependence of humans on their environment and the services provided by ecosystems portends a new and expanded focus on understanding ecological patterns and processes across geographic (e.g., continents) and political (e.g., country) boundaries. Such a focus, points to the need for many future advances in how data are managed for long term ecological research. From both a technical and sociocultural perspective, we postulate that some of the principal information management tools needed for the next decade will include:

- a new class of user-friendly metadata management tools and standards that better enable understanding, use and re-use of data such as processes built into sensors and instruments that automatically capture metadata at the time of data generation;

- semantic mediation tools and approaches that facilitate data integration across multiple spatiotemporal scales and levels of biological and physical organization, including the automated encapsulation of data provenance;
- analysis and visualization approaches that support reasoning about and representing scientific uncertainty;
- tools that promote trust of information products such as automated quality assessment of data products/service providers, development of recommender systems for scientific data, as well as development of security and identity-authentication systems that work world-wide;
- reward systems that promote recognition of good science practices through data citation and data use statistics; and
- downloadable and user-oriented videos, course modules, and best practices that engage and educate scientists, students, and the public.

Acknowledgments

The work reported in this paper was supported through awards from the U.S. National Science Foundation (NSF 0621014, 0620482). William Michener would like to acknowledge additional financial support provided from NSF (#0753138 and #0830944) as well as technical support provided by Kayla Achen.

References

Ailamaki, A., Kantere, V., Dash, D., 2010. Managing scientific data. Communications of the ACM. 53 (6), 67–78.

Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M., 2009. Some simple guidelines for effective data management. ESA Bulletin 90, 205–214.

Brunt, J., McCartney, P., Baker, K., Stafford, S., 2002. The future of ecoinformatics in long term ecological research. In: Callaos, N., Porter, J., Rishe, N. (Eds.), Proceedings of 2002 Systemics, Cybernetics, and Informatics Symposium. July 14–18, 2002 Orlando, Florida, pp. 367–372.

Callahan, J.T., 1984. Long-term ecological research. BioScience 34, 363–367.

Collins, S.L., Bettencourt, L.M.A., Hagberg, A., Brown, R.F., Moore, D.I., Bonito, G., Delin, K.A., Jackson, S.P., Johnson, D.W., Burleigh, S.C., Woodrow, R.R., McAuley, J.M., 2006. New opportunities in ecological sensing using wireless sensor networks. Frontiers in Ecology 4, 402–407.

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (eml) and principles for metadata creation. Bulletin of Ecological Society of America 86, 158–168.

Franklin, J.F., Bledsoe, C.S., Callahan, J.T., 1990. Contributions of the long-term ecological research program. BioScience 40, 509–523.

Gosz, J.R., Waide, R.B., Magnuson, J.J., 2010. Twenty-eight years of the US-LTER program: experience, results, and research questions. Long-Term Ecological Research 59–74.

Henshaw, D.L., Sheldon, W.M., Remillard, S.M., Kotwica, K., 2006. ClimDB/HydroDB: a web harvester and data warehouse approach to building a cross-site climate and hydrology database. Proceedings of the 7th International Conference on Hydroscience and Engineering (ICHE 2006). Michael Piasecki and College of Engineering. Drexel University, Philadelphia, USA.

Higgins, D., Berkley, C., Jones, M.B., 2002. Managing heterogeneous ecological data using Morpho. 14th International Conference on Scientific and Statistical Database Management, Edinburgh, Scotland, July 24–26, 2002, p. 69.

Hobbie, J.E., Carpenter, S.R., Grimm, N.B., Gosz, J.R., Seastedt, T.R., 2003. The US long term ecological research program. BioScience 53, 21–32.

Keller, M., Schimel, D.S., Hargrove, W.W., Hoffman, F.M., 2008. A continental strategy for the National Ecological Observatory Network. Frontiers in Ecology and the Environment 6 (5), 282–284.

Kortz, M., Conners, J., Yarmey, L., Baker, K.S., 2009. LTER community resources: unit dictionary and unit registry [abstract]. American Geophysical Union Fall Meeting, Dec. 14–18, 2009, San Francisco, CA. Abstract #IN21B-1056.

Kratz, T.K., Magnuson, J.J., Bayley, P., Benson, B.J., Berish, C.W., Bledsoe, C.S., et al., 1995. Temporal and spatial variability as neglected ecosystem properties: lessons learned from 12 North American ecosystems. In: Rapport, D., Calow, P. (Eds.), Evaluating and Monitoring the Health of Large-Scale Ecosystems. Springer, New York, pp. 359–383.

Kratz, T.K., Arzberger, P., Benson, B.J., Chiu, C.Y., Chiu, K., Ding, L., Fountain, T., Hamilton, D., Hanson, P.C., Hu, Y.H., Lin, F.P., McMullen, D.F., Tilak, S., Wu, C., 2006. Toward a Global Lake Ecological Observatory Network, vol. 145. Publications of the Karelian Institute, pp. 51–63.

Laney, C.M., Peters, D.P.C., 2006. EcoTrends in long-term ecological data: a collaborative synthesis project, introduction and update. LTER DataBits Spring 2006. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/ 2006.

Lin, C.C., Porter, J.H., Hsiao, C.W., Lu, S.S., Jeng, M.R., 2008. Establishing an EML-based data management system for automating analysis of field sensor data. Taiwan Journal of Forest Science 23, 279–285.

Magnuson, J.J., 1990. Long-term ecological research and the invisible present. BioScience 40, 495–501.

McCartney, P., Jones, M., 2002. Using XML-encoded metadata as a basis for advanced information systems for ecological research. In: Callaos, N., Porter, J., Rishe, N. (Eds.), Proceedings of 2002 Systemics, Cybernetics, and Informatics Symposium. July 14–18, 2002 Orlando, Florida, pp. 379–384.

Michener, W.K., 1986. Research Data Management in the Ecological Sciences. University of South Carolina Press, Columbia, SC.

Michener, W.K., Nottrott, R., 1990. LTER Core Data Set Catalog. LTER Network Office, Seattle, WA.

Michener, W.K., Waide, R.B., 2009. The evolution of collaboration in ecology: lessons from the United States Long Term Ecological Research Program. In: Olson, G.M., Zimmerman, A., Bos, N. (Eds.), Scientific Collaboration on the Internet. MIT Press, Boston, pp. 297–310.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G., 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7, 330–342.

Nottrott, R., Jones, M.B., Schildhauer, M., 1999. Using XML-structured metadata to automate quality assurance processing for ecological data. Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD, April 6–7, 1999.

Porter, J.H., 2000. Scientific databases. In: Michener, W.K., Brunt, J. (Eds.), Ecological Data: Design, Processing and Management. Blackwell Science Ltd., London.

Porter, J., 2010. A controlled vocabulary for LTER datasets. LTER Databits: Information Management Newsletter of the LTER Network. http://databits.lternet.edu/node/105 <Accessed 5/5/2010>.

Porter, J., Arzberger, P., Braun, H., Bryant, P., Gage, S., Hansen, T., Hanson, P., Lin, C., Lin, F., Kratz, T., Michener, W., Shapiro, S., Williams, T., 2005. Wireless sensor networks for ecology. BioScience 55 (7), 561–572.

Porter, J., Nagy, E., Hanson, P.C., Kratz, T.K., Collins, S., Arzberger, P.A., 2009. New eyes on the world: advanced sensors for ecology. BioScience 59, 385–397.

Servilla, M., Brunt, J., San Gil, I., Costa, D., 2006. PASTA: a network-level architecture design for generating synthetic data products in the LTER Network. LTER DataBits Fall 2006. http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/ 2006.

Servilla, M., Costa, D., Laney, C., San Gil, I., Brunt, J., 2008. The EcoTrends web portal: an architecture for data discovery and exploration. Proceedings of the Environmental Information Management Conference 2008, September 10–11, 2008, Albuquerque, New Mexico, pp. 139–144.

Sheldon, W.M., 2002. GCE data toolbox for MATLAB — Software tools for metadata-based analysis, visualization and transformation of ecological data sets. http://gcelter.marsci.uga.edu/public/im/tools/data_toolbox.htm 2002.

Sheldon Jr., W.M., 2008. Dynamic, rule-based quality control framework for real-time sensor data. In: Gries, C., Jones, M.B. (Eds.), Proceeding of the Environmental Information Management Conference 2008 (EIM 2008): Sensor Networks. Albuquerque, New Mexico, pp. 145–150.

Swanson, F.J., Sparks, R.E., 1990. Long-term ecological research and the invisible place. BioScience 40, 502–508.

U.S. Long Term Ecological Research Network (LTER). 2007. The Decadal Plan for LTER: Integrative Science for Society and the Environment. LTER Network Office Publication Series No. 24, Albuquerque, New Mexico. 154 pages.

Vanderbilt, K., Michener, W.K., 2007. Information management standards and strategies for net primary production data. In: Fahey, T., Knapp, A. (Eds.), Principles and Standards for Measuring Primary Production. Oxford University Press, New York, pp. 12–26.