LTER INFORMATION MANAGERS SKILLSET AND TRAINING RESOURCES WORKING GROUP RECOMMENDATIONS

request for review by IM Exec: 2017-03-15
request for review by IMC: 2017-05-03

SECTION 1. BACKGROUND AND SCOPE

Information management is an integral component of an LTER site, facilitating the flow of scientific information from conception through publication, and providing critical services to support internal site management and function. An Information Management System (IMS) is comprised of hardware, software, and people resources needed to support the generation and collection of scientific information, and to make those resources available to the LTER and broader scientific communities. An IMS is by nature multi-faceted and comprised of multiple components (e.g., databases; web-, database-, and file-servers), and an Information Manager must possess the knowledge and skills requisite to develop, coordinate, and maintain these inter-connected components.

Although an Information Manager must possess a perfunctory level of knowledge related to the structure and function of all components that contribute to an LTER site IMS, the scope and degree of knowledge required by the Information Manager will vary depending on the level of institutional support for underlying components, and the type of IMS employed. For example, a host institution may provide file servers available for use by the LTER site such that an Information Manager must have the knowledge to be able to interact with the file servers but not necessarily to create or maintain them. Given the diversity of possible scenarios and that the incoming Information Manager is likely to have pre-existing technical knowledge, this document will focus on skills that are cross-cutting among LTER Information Managers regardless of institutional resources or the specific configuration of the site IMS.

SECTION 2. RECOMMENDATIONS FOR NEW LTER SITES AND INFORMATION MANAGERS: A SYSTEMS-LEVEL AND LTER INFORMATION MANAGEMENT COMMUNITY-CENTRIC APPROACH

The committee recognizes two realities pertaining to training Information Managers at, particularly, new LTER sites and, to a lesser extent, new Information Managers coming into existing LTER sites. First, the most formidable obstacle to learning or developing an IMS is not the application of any single component, but rather the integration of subcomponents into a holistic, functional IMS. Second, the most valuable resource for gaining insight into the development and operation of a successful IMS is the community of LTER Information Managers. In light of these, the committee recommends foremost a systems-level approach to training in which new Information Managers would gain insight into IMS implementation via direct interaction and

hands-on training with members of the current community of Information Managers, followed by or with the option for focused training of particular skills provided by the Environmental Data Initiative (EDI) or other entity that provides information management training.

2.A. Advisory panel and mentorship team

The committee recommends the formation of a panel consisting of current LTER Information Managers who would provide an overview of the IMS employed at their respective sites in a virtual or in-person conference with incoming Information Managers. The panel serves several purposes. First, it provides system-level overviews of IMS implementation of varying flavors, which yields insight into how individual technological components may be linked to form an overarching IMS. This approach is critical to provide the context for more focused training. For example, focused training may provide instruction on the use of Metabase as an information resource for a new LTER site. However, without background or context, it may not be apparent to an incoming Information Manager as to how or why Metabase may be employed, particularly relative to alternative approaches. While the committee feels it is important to present multiple approaches to IMS implementation, the need to reduce (or, at least, not contribute further to) far too many idiosyncratic approaches to IMS in practice across the LTER network is recognized. As such, system integration should be the primary focus of the panel whereas training on selected IMS subcomponents (e.g., R for EML production) should rely on training developed for wider adoption by EDI. Second, rapport among panel members and incoming Information Managers provides a connection through which the incoming Information Managers should feel comfortable reaching out as questions inevitably arise. Such a connection is particularly critical in light of the now fewer opportunities available to the LTER Information Management Committee (IMC) for in-person meetings, which previously were invaluable educational opportunities for new Information Managers. Third, an advising panel and mentorship team is an approach that would most effectively and efficiently provide instruction and insight as to the roles, responsibilities, and operation of new Information Managers within the LTER IMC, and regarding interaction with the LTER NCO and network. Finally, the committee suggests that a panel and mentor approach is the most effective means to address the nuances of LTER information management. What are, for example, the most effective approaches to soliciting research data and metadata from graduating students? Such insight comes from experience, and the value of an experienced Information Manager passing on such wisdom cannot be understated.

2.B. On-site mentoring program

As a follow-up to the advising panel and mentorship team, and in conjunction with focused training by EDI or other entity, the committee recommends an on-site mentoring program in which new Information Managers have the opportunity to visit an LTER site where that site maintains an IMS of particular interest or relevance to an incoming Information Manager. The goal of this program is to provide for a more

in-depth overview of the IMS at that site with hands-on training. This program is contingent upon the availability and interest of a site and mentor (i.e., site Information Manager) to serve in this host capacity as a structured session in which key components of the IMS are detailed would be essential to facilitate sufficient knowledge transfer during a relatively brief visit. Two funding options for the visiting Information Manager are envisaged: (1) given the anticipated benefit of the on-site mentoring program to the incoming Information Manager, it is expected that travel costs will be an acceptable budget expense to a new (or existing) LTER site; (2) it may be feasible to cover travel costs through an award from EDI (proposal required).

SECTION 3. RECOMMENDATIONS: REQUISITE SKILLS AND FOCUSED TRAINING FOR LTER INFORMATION MANAGERS

Although the committee stresses the importance of systems-level and LTER Information Management community-centric training to provide the necessary background and context for site IMS implementation, it is recognized that all LTER Information Managers require a certain set of skills, and that focused training in these areas is equally warranted. Following is a list and brief overview of focused training that the committee recommends the EDI should develop and implement for the benefit of incoming and, in some cases, current LTER Information Managers.

3.A. Generate and employ Ecological Metadata Language (EML) metadata

All LTER Information Managers must possess the knowledge to generate EML for the data that they manage. At a minimum, this means that the Information Manager knows which EML elements are required metadata, and understands the EML schema sufficiently to manually edit a EML document and ensure that it is well formed and compliant with the schema. To address these needs, an Information Manager must have mastery of a tool or tools to create EML. Several tools are available, and Information Managers should learn to use the R EML (https://ropensci.github.io/EML/) Package as recommended by EDI and/or Morpho (or its pending successor) as these tools stand alone and, thus, do not rely on the existence of a relational database or custom scripts to generate EML. In addition, the GCE Toolbox is another stand-alone tool for generating complete PASTA-compliant EML, is particularly effective for streaming sensor data management, and is a valuable tool of the information workflow at many LTER sites.

New Information Managers should also be made aware of how EML can be leveraged to produce web content through the application of stylesheets.

Table 3.A. EDI and the LTER Information Managers should co-develop training modules to cover the following topics related to EML

topic	subtopic	content providers
EML schema		LTER
XML Editors		LTER
Overview of current approaches for generating EML	GCE Toolbox	EDI
	DEIMS	
	Metabase Type Systems (database to EML scripts)	
	Morpho (or its pending successor)	
	R EML Package	
	Spreadsheet translators	
EML generation best practices (http://im.lternet.edu/node/910)	Construction of Scope IDs	LTER
	ORCID IDs	
	File naming conventions	
	Clarifying the roles of originators, contacts, associated parties, etc.	
	EML data types (table, Spatial Vector, Spatial Raster, Software, otherEnity)	
	When to use the Other Entity (otherEntity) type	
	Unit Dictionary, STMML (scientific-technical-medical markup	
	Keywords (controlled vocabulary)	
	Access Rules (metadata, data)	
	Access Rules (metadata, data)	
	Provenance	
	Attribute typing (ordinal, nominal, interval, ratio)	
	Titles (e.g., constructing for a global (not local) audience), abstracts, methods	
Using EML at a site	XSLT (Extensible Stylesheet Language Transformations)	LTER
	Present EML metadata on a website	

3.B. Interfacing with the Data Portal and PASTA+

Interfacing with the Data Portal and PASTA+ is a fundamental task central to the responsibilities of all LTER Information Managers. The committee recommends training on any and all aspects of interfacing with the Data Portal and PASTA+, such as, but not necessarily limited to, loading and extracting data and metadata, the web interface, the API, the basic workings of the service (e.g., which EML elements are extracted for inclusion in a data package citation), versioning, etc.

Table 3.B. EDI should develop training on the following topics as related to *Interfacing with the Data Portal and PASTA+*

topic	subtopic	content providers
An introduction to the Data Portal and PASTA+	Upload data and metadata to the Data Portal	EDI
	The Quality Checker	
The PASTA+ API	What you can do with it	EDI
	How to use the PASTA API (start here: http://im.lternet.edu/node/1280)	
	Extract a report of data sets with Digital Object Identifiers (DOIs) for a LTER site	
	Track data downloads and metadata views	

3.C. Develop and maintain a local catalog of site-specific data

Making site-specific data available to the broader scientific community in a systematic, formalized, and structured way such that data are well described and readily discoverable is a central tenet of LTER information management. Addressing this central function requires the development and maintenance of a site-based catalog of site-specific data products. This task is addressed using a variety of technologies across the LTER network, but has historically included three broad approaches: (1) an XML/EML-generated catalog, (2) a database-generated catalog, and (3) a static, manually constructed catalog. The landscape, however, is changing with the universal adoption and use of the the (LTER/EDI) Data Portal. The expectation that data products are to be housed in the Data Portal eliminates the need for an LTER site to construct a separate cataloging system to display, and make discoverable and available site-specific data. Although a site-based data catalog is required, the more efficient and now recommended approach to constructing this catalog is to pipe the datasets specific to that site from the Data Portal. In conjunction with the training recommendations outlined for *Interfacing with the Data Portal and PASTA+*, it is recommended that

Information Managers acquire sufficient training with the PASTA+ API and with web technologies to facilitate interfacing with the Data Portal to generate a site-based data catalog on the site website.

Table 3.C. The LTER Information Management Committee and EDI should co-develop training on the following topics as related to developing and maintaining a local catalog of site-specific data

topic	content providers
Extracting information from PASTA (see Table 3.B. PASTA+ API above)	
Styling the rendering of results (see Table 3.A. Using EML at a site above)	

3.D. Database design, administration, security, and implementation

Although the type, implementation, and use of Relational Database Management Systems (RDBMS) varies across the LTER network, that an RDBMS will be employed to meet some or many site information management needs is likely if not certain. Though the topic is broad, it is nonetheless imperative that incoming Information Managers have the basic skills needed to design, create, and administer (address permissions, perform at least basic Structured Query Language (SQL) operations, and backup) an RDBMS. In light of the large selection of RDBMS software solutions, the near limitless implementations and uses of an RDBMS, and the plethora of existing RDBMS-related training resources, the committee recommends that incoming Information Managers receive RDBMS training developed by EDI that is tailored to likely applications of an RDBMS at a LTER site.

Table 3.D. EDI should develop training on the following topics as related to the application of databases to the IMS at LTER sites

topic	subtopic	content providers
LTER database applications	data inventory schema (e.g., tracking data submission, status, publication details)	EDI
	metadata schema	
	project management/tracking schema	
Database introductory materials	overview of database design	EDI
	SQL	
	database backup	
Advanced database topics	stored procedures	EDI
	triggers	
	overview of NoSQL approaches	
	Programmatic interfaces to databases (Python, Ruby, PHP, PERL, JAVA, R)	

3.E. Quality control to ensure data integrity

Although it is the responsibility of the data provider to ensure that the data submitted is accurate and that the measurements are appropriate, the Information Manager may employ various checks to verify and, possibly, improve the overall quality or trustworthiness of the data. The Information Manager may, for example, ensure that all reported study units are accounted for (e.g., that data from all expected study plots are included), data entry errors are caught (e.g., duplicate observations), use of codes is uniform throughout the database, and the data values are realistic (e.g., 1000 kg mice were not sampled). Systematic errors introduced into the data set by sensors, such as sensor drift, can also be detected then noted or corrected. The integrity of the data file itself can be monitored using checksums to determine if errors or changes have been introduced in transmitted data. New Information Managers should be made aware of these different types of checks, and the committee recommends that examples of how to implement them should be addressed by the *Advisory Panel and Mentorship Team*.

Table 3.E. The *Advisory Panel and Mentorship Team* should touch on the following topics as related to quality control, and appropriate training modules developed by the LTER IM Community

topic	**content providers**
Standard quality checks: range, duplication, domain, data type, spikes, jumps, logical consistency	LTER
File data integrity: checksums, length	LTER

3.F. Lineage: generate a record of any and all modifications to data

The journey from data collection through publication may be a long one, and maintaining a record of any modifications to data along the way is critical to ensuring the integrity and reproducibility of the research. Original and intermediate products should be preserved, along with code and notes regarding changes to the data. Git, GitHub, and other code-management systems may be usefully applied to track changes, but they may have limitations with respect to longevity. Code-based and database approaches, where original data are maintained unchanged and any changes are made using programs can also be used.

Table 3.F. The *Advisory Panel and Mentorship Team* should touch on the following topics as related to data lineage, and EDI should develop appropriate training modules

topic	content providers
Versioning techniques	LTER
History tables	LTER
Logging	LTER
Good coding practices (documenting your code)	LTER
README files	LTER
Knowledge base of known issues and how they were resolved	LTER
Provenance-aware project management systems (e.g., OSF.io)	LTER

Additionally, Information Managers should consider resources such as Software Carpentry for training on tools such as Git and GitHub.

3.G. Acquire metadata (and data)

The core mission of making site-specific data available to the broader scientific community begins with first acquiring the data and associated metadata. Obtaining the latter can be particularly difficult owing to effort required on the part of the investigator or data generator to provide metadata such that the data product is sufficiently described. Numerous approaches are employed across the LTER network to facilitate the transfer or metadata from the investigator or data generator to the Information Manager, including structured spreadsheets, text or online forms, or simply by word-of-mouth (not recommended). It is expected that new Information Managers will gain insight into effective approaches to acquiring metadata from the aforementioned community-centric training (e.g., by the Advisory Panel and Mentorship Team). In addition, the committee recommends that EDI take up the task of identifying a recommended approach to acquiring metadata that is applicable to any environment (e.g., not dependent on a DEIMS implementation), and providing resources toward formalizing and developing it, and constructing corresponding training modules.

3.H. Security

Security of users, data, and systems is a key element of any successful system. At some sites, security, or certain aspects of it, may be handled by a systems administrator, and, historically, those roles have overlapped. However, with constantly escalating threats, it is critical that Information Managers have some basic knowledge of security techniques regardless the level of systems support provided by the host institution.

Table 3.H. EDI should develop training on the following topics as related to systems security

topic	**subtopic**	**content providers**
System security	Antivirus	EDI
	Updates	
	Backups (off site, tested)	
Network security	Web security	EDI
	Firewalls	
	Encryption/Certificates/SSL	
	Email Screening	
Database security	Host limits	EDI
	Firewalls	
	Backups	
User security	Good passwords	EDI