Int. J. Metadata, Semantics and Ontologies, Vol.

The Long-Term Ecological Research community metadata standardisation project: a progress report

Inigo San Gil*

Department of Biology,
University of New Mexico,
LTER Network Office,
MSC03 2020, Albuquerque, NM 87131, USA
E-mail: isangil@lternet.edu
*Corresponding author

Karen Baker

Scripps Institution of Oceanography,
University of California at San Diego,
La Jolla, CA 92093, USA
E-mail: kbaker@ucsd.edu

John Campbell and Ellen G. Denny

Hubbard Brook LTER,
USDA Forest Service,
271 Mast Road, Durham, NH 03824, USA
E-mail: jlcampbell@fs.fed.us
E-mail: ellen.denny@yale.edu

Kristin Vanderbilt

Department of Biology,
University of New Mexico,
Sevilleta LTER, MSC03 2020,
Albuquerque, NM 87131, USA
E-mail: vanderbi@sevilleta.unm.edu

Brian Riordan, Rebecca Koskela and Jason Downing

Bonanza Creek Long Term Ecological Research Station,
University of Alaska Fairbanks,
909 N. Koyukuk, Fairbanks, AK 99709, USA
E-mail: brian.riordan@gmail.com
E-mail: fnrjk@uaf.edu
E-mail: jpdowning@alaska.edu

Sabine Grabner

University of Applied Sciences,
Hochschulstrasse 1 E210 A-6850, Dornbirn, Austria
E-mail: sabine.grabner@gmail.com

Eda Melendez

Luquillo Experimental Forest LTER,
University of Puerto Rico at Rio Piedras,
San Juan, PR 00931, USA
E-mail: emelendez@lternet.edu

Jonathan M. Walsh

Baltimore Ecosystem Study LTER,
Cary Institute of Ecosystem Studies,
2801 Sharon Turnpike, Millbrook, NY 12545, USA
E-mail: walshj@caryinstitute.org

Mason Kortz, James Conners and Lynn Yarmey

Scripps Institution of Oceanography,
University of California at San Diego,
La Jolla, CA 92093, USA
E-mail: mkortz@ucsd.edu
E-mail: jconners@ucsd.edu
E-mail: yarmey@coast.ucsd.edu

Nicole Kaplan

Short Grass Steppe LTER,
Colorado State University,
Fort Collins, CO 80523, USA
E-mail: nkaplan@lternet.edu

Emery R. Boose

Harvard Forest LTER,
Harvard University,
Petersham, MA 01366, USA
E-mail: boose@fas.harvard.edu

Linda Powell

Florida Coastal Everglades LTER,
Florida International University,
University Park, Miami, FL 33199, USA
E-mail: Powell@fiu.edu

Corinna Gries and Robin Schroeder

Central Arizona Phoenix LTER,
Global Institute of Sustainability,
Arizona State University,
Tempe, AZ 85287, USA
E-mail: cgries@asu.edu
E-mail: pmccartn@nsf.gov
E-mail: Robin.Schroeder@asu.edu

Todd Ackerman

Niwot Ridge LTER,
University of Colorado,
1560 30th Street, Boulder, CO 80309-0450, USA
E-mail: todda@colorado.edu

Ken Ramsey

Jornada del Muerto LTER,
Department of Biology,
New Mexico State University,
Jornada LTER, Las Cruces, NM 88003, USA
E-mail: kramsey@jornada.nmsu.edu

Barbara Benson and Jonathan Chipman

Center for Limnology,
University of Wisconsin-Madison,
North Temperate Lakes LTER,
680 N. Park Street Madison, WI 53706, USA
E-mail: bjbenson@wisc.edu
E-mail: jchipman@wisc.edu

James Laundre

Marine Biological Laboratory Ecosystems,
Arctic LTER, Center Woods Hole MA 02543, USA
E-mail: jiml@mbl.edu

Hap Garritt

Plum Island Ecosystem LTER,
Marine Biological Laboratory Ecosystems,
Center Woods Hole MA 02543, USA
E-mail: hgarritt@mbl.edu

Don Henshaw

Andrews Forest LTER,
USDA Forest Service Pacific NW Research Station,
3200 SW Jefferson Way Corvallis, OR 97331, USA
E-mail: don.henshaw@oregonstate.edu

Barrie Collins

Anthropology Department,
Coweeta LTER, Ecolaboratory,
251 Baldwin Hall University of Georgia Athens,
GA 30602, USA
E-mail: barriec@uga.edu

Christopher Gardner

McMurdo Dry Valleys LTER,
Byrd Polar Research Center,
1090 Carmack Rd Columbus, OH 43210, USA
E-mail: gardner.177@osu.edu

Sven Bohm

Kellog Biological Station,
Department of Crop and Soil Sciences,
Michigan State University East Lansing,
MI 48824, USA
E-mail: bohms@msu.edu

Margaret O'Brien

Santa Barbara Coastal LTER,
Marine Science Institute,
University of California at Santa Barbara,
Santa Barbara, CA 93116, USA
E-mail: mob@msi.ucsb.edu

Jincheng Gao

Konza Prairie LTER,
Division of Biology,
Kansas State University 116 Ackert Hall,
Manhattan, KS 66506, USA
E-mail: jcgao@ksu.edu

Wade Sheldon

Georgia Coastal Ecosystems LTER,
Department of Marine Sciences,
University of Georgia,
Athens, GA 30602, USA
E-mail: Sheldon@uga.edu

Stephanie Lyon and Dan Bahauddin

Cedar Creek Natural History Area LTER,
100 Ecology Building,
1987 Upper Buford Circle,
St. Paul, MN 55108, USA
E-mail: pimmx001@umn.edu
E-mail: danbaha@umn.edu

Mark Servilla, Duane Costa and James Brunt

LTER Network Office,
Department of Biology,
University of New Mexico,
MSC03 2020, Albuquerque, NM 87131, USA
E-mail: servilla@lternet.edu
E-mail: dcosta@lternet.edu
E-mail: jbrunt@lternet.edu

Abstract: We describe the process by which the Long-Term Ecological Research (LTER) Network standardised their metadata through the adoption of the Ecological Metadata Language (EML). We describe the strategies developed to improve motivation and to complement the information technology resources available at the LTER sites. EML implementation is presented as a mapping process that was accomplished per site in stages, with metadata quality ranging from 'discovery level' to rich-content level over time. As of publication, over 6000 rich-content standardised records have been published using EML, potentially enabling the goal of machine-mediated, metadata-driven data synthesis.

Keywords: metadata; metadata management; standardisation; EML; ecological metadata language.

Reference to this paper should be made as follows: San Gil, I., Baker, K., Campbell, J., Denny, E.G., Vanderbilt, K., Riordan, B., Koskela, R., Downing, J., Grabner, S., Melendez, E., Walsh, J.M., Kortz, M., Conners, J., Yarmey, L., Kaplan, N., Boose, E.R., Powell, L., Gries, C., Schroeder, R., Ackerman, T., Ramsey, K., Benson, B., Chipman, J., Laundre, J., Garritt, H., Henshaw, D., Collins, B., Gardner, C., Bohm, S., O'Brien, M., Gao, J., Sheldon, W., Lyon, S., Bahauddin, D., Servilla, M., Costa, D. and Brunt, J. (2009) 'The Long-Term Ecological Research community metadata standardisation project: a progress report', *Int. J. Metadata Semantics and Ontologies*, Vol.

Biographical notes: I. San Gil received his PhD in Mechanical Engineering from Yale University in 2001. He is currently the metadata project coordinator and senior systems analyst for the National Biological Information Infrastructure and Long-Term Ecological Network. His current research interests include metadata management systems, bioinformatics, and metadata-driven systems.

Karen Baker holds an MS from the University of California at Los Angeles, and she is currently the information manager at Scripps Institution of Oceanography for CalCOFI, Palmer LTER and California Current Ecosystem LTERs.

John Campbell holds a PhD from the State University of New York and he is currently the information manager at Hubbard Brook LTER.

Ellen G. Denny holds a MFS from the Yale School of Forestry and Environmental Studies, and is part of the information management team for the Hubbard Brook LTER site.

Kristin Vanderbilt received her PhD (Biology) at the U. of New Mexico, where she is currently an Associate Research Professor and the Sevilleta LTER Information Manager.

Brian Riordan received an MS from the University of Alaska – Fairbanks, he currently works at the private sector on GIS.

Rebecca Koskela is a Bioinformatics Specialist at the Arctic Region Supercomputing Center at the University of Alaska Fairbanks campus, Rebecca was a member of the senior management team at the Aventis Cambridge Genome Center.

Jason Downing is the current Information Manager at the Bonanza Creek LTER.

Sabine Grabner received her MS (Meteorology) from the University of Innsbruck, Austria. She was the Information Manager at the Moorea Coral Reef LTER and she is currently employed at the University of Applied Sciences in Dornbirn, Austria.

Eda Melendez holds an MS (Applied Mathematics) from Rutgers University, New Jersey, USA.

Jonathan M. Walsh is currently the information manager at the Baltimore Ecosystems LTER.

Mason Kortz is a staff member, a student and a programmer at the Scripps Institution of Oceanography.

James Conners is a staff member, a student and a programmer at the Scripps Institution of Oceanography.

Lynn Yarmey is a staff member, a student and a programmer at the Scripps Institution of Oceanography.

Nicole Kaplan is the Short Grass Steppe LTER information manager. Nicole was previously a field technician at the SGS LTER site.

Emery R. Boose received his PhD in 1988 from Harvard University (Sanskrit & Indian Studies). He is the information manager at the Harvard Forest LTER.

Linda Powell has a Master's Degree in Geology and he is currently the information manager at the Florida Coastal Everglades LTER.

Corinna Gries is an Associate Research Professor at the Arizona State University and coordinates the information management tasks at the Arizona-Phoenix LTER.

Robin Schroeder was a staff member of the Central Arizona Phoenix LTER, Global Institute of Sustainability information management team.

Todd Ackerman is the Niwot Ridge LTER Information Manager University of Colorado.

Ken Ramsey is the Jornada del Muerto LTER Information Manager.

Barbara Benson is the North Temperates Lakes LTER Information Manager. Barbara also was the LTER Information Management Committee Chair.

Jonathan Chipman was NTL's GIS information management leader at the Center for Limnology.

James Laundre is the Arctic LTER Information Manager at the Marine Biological Laboratory Ecosystems (Woods Hole, MA).

Hap Garritt is the Plum Island Ecosystem LTER Information Manager at the Marine Biological Laboratory Ecosystems (Woods Hole, MA).

Don Henshaw is the Andrews Forest LTER information management team leader and has chaired numerous roles at the LTER governing bodies.

Barrie Collins is the Coweeta LTER Information Manager and comes from the pricate sector, where he worked in GIS systems. He has participated in the Information Management Executive Committee at LTER.

Christopher Gardner holds an MS from Ohio State in Geological Science and he is currently the Information Manager at McMurdo Dry Valleys LTER.

Sven Bohm is currently the Kellogg Biological Station LTER – Information Manager.

Margaret O'Brien is currently the Information Manager at the Santa Barbara Coastal LTER.

Jincheng Gao is the Konza Prairie LTER information manager.

Wade Sheldon holds an MS, University of Georgia and he is currently the information manager at the Georgia Coastal Ecosystems LTER.

Stephanie Lyon holds a BS Degree (Biology) from the UMN, she was the Information Manager at Cedar Creek Natural History Area LTER until 2007.

Dan Bahauddin holds a BS from the University of Minnesota Aerospace Engineering Department and he is currently the information manager at the Cedar Creek LTER.

Mark Servilla is the Lead Scientist for the Network Information System (NIS) at LTER. Prior to his current position at LTER Network Office, Mark worked in the private sector at Photon Research Associates (PRA), Inc. He holds graduate degrees in Earth and Planetary Sciences (Volcanology) and Computer Science, both from the University of New Mexico. As a post-doctoral Research Associate with the Alaska Volcano Observatory (AVO) at the University of Alaska, Fairbanks.

Duane Costa is Analyst/Programmer for the Network Information Systems. He holds a MS in Computer Science at the University of Rhode Island. Before joining LTER, he worked for ABAQUS, Inc.

James Brunt is an Associate Director for Information Management of the LTER Network Office, he leads and supervises a staff of six who provide operations and maintenance of LTER cyberinfrastructure, design and develop the LTER Network Information System, and provide stewardship of LTER Network databases and websites. He pursued a unique MS mixing Ecology, Computer Science, and Experimental Statistics at NMSU.

1 Introduction

Is the petabyte information age (Anderson, 2008) overwhelming the scientific community? Are scientists, and the information managers, who assist them, ready and able to produce, manage and preserve the wealth of data that can now be collected, distributed and analysed anywhere by anyone?

Data retains its value over time only as long as it is well documented. Data documentation is referred to as metadata, or data about data. While some minimum level of metadata is critical for a data set to be usable, extensive and high-quality metadata facilitates scientific analysis of complex heterogeneous datasets. The LTER Network identified the need to standardise metadata and selected

the EML as the vehicle to achieve standardisation. EML is an XML schema-based, community-driven language that describes ecological data. By the summer of 2007 (San Gil, 2007) all 26 LTER sites had achieved some degree of metadata standardisation. This feat so far amounts to about 6000 metadata records. This paper describes and analyses the process of how the LTER achieved metadata standardisation.

We start this paper with an overview of the LTER Network and its mission, and explain how the need for metadata standardisation developed within and across LTER sites. We provide the background of how the EML was adopted as the official language of the LTER Network. We include an overview of what the EML: is and how it meets the needs of the LTER network (a reader unfamiliar with LTER and EML may want to consult the references provided therein). We then describe the diverse strategies and methods employed to implement the EML standard in specific LTER sites. We point out the advantages of different standardisation methods and how these methods cater to each site's data management procedures, resources and capabilities. We conclude by highlighting the achievements and the lessons learned in the standardisation process, and then outlining future steps to complete and improve upon the standardisation of the metadata.

2 The LTER in a nutshell

The LTER Network (Lternet website, 2008) was officially founded in 1980. Initially, six sites distributed across the USA were commissioned by the US Congress to make continuous studies of the environment. Today, the LTER encompasses 26 sites and is an established National Science Foundation (NSF) programme in biology. The sites cover diverse ecosystems across the continental US and extend to Alaska, the Caribbean, the Pacific and Antarctica. Over 2000 scientists are using these sites to perform long-term interdisciplinary research on diverse ecosystems. These LTER sites comprise most of the planetary habitats including estuaries, lakes, oceans, coral reefs, deserts, prairies, and forests, alpine and Arctic tundra. LTER sites also perform studies in urban areas and production agriculture. The LTER mission is to provide the scientific community, policy makers, and society with the knowledge and predictive understanding necessary to conserve, protect, and manage the nation's ecosystems, their biodiversity, and the services they provide. There are over 40 other nations which have formed LTER networks under the umbrella of a global network called International LTER (ILTER).

While the LTER is a network, all sites have a high degree of autonomy and manage their own research and data. Sites are managed in partnership with universities or with the US Forest Service. All the 26 LTER sites have at least one information manager. The LTER Network also provides information technology related services through a centralised office.

Long-term research generates massive amounts of data. Consider, on just one site, the continuing experiments spanning decades, with new studies commencing every year, with different scientists and new data managers rotating in to manage the site and the data. Also, consider the rapidly changing technology that, since the 1980s has exponentially magnified the capacity to collect, store and analyse data. Multiply this one site by 26 sites, all with their own personality and history, and then consider the developing need to share and compare data across sites. Over time the need for standardised metadata became obvious and the challenge then became how to implement these standards across the diverse sites.

Responding to the aforementioned challenge, after lengthy discussions (Harmon, 2003) in 2003 the LTER governing body officially approved the adoption of the EML as the standard for the LTER network metadata.

3 The Ecological Metadata Language

The EML is used to describe scientific metadata in detail. The EML is implemented in an XML schema, making the resulting metadata records independent of any computer platform or information system. EML inherits all benefits of XML, including the inherent ability to leverage many Service-Oriented Architectures (SOAs) that are driving today's information exchanges over the internet. EML is implemented as a suite of extensible and reusable modules, such as a resource module that accommodates basic geotemporal references, data identification, and provenance information. It also has modules that describe in detail the many physical and logical aspects of the data entities being described.

EML was born at the National Center for Ecological Analysis and Synthesis (NCEAS), where the early EML versions were originally created based on a paper by Michener et al. (1997). Papers by McCartney and Jones (2002) and references therein offer further insight on EML as the facilitating tool to build a federated network of data.

EML was conceived to provide a platform for metadata-driven data analysis. The EML founders (See eml-dev) leveraged work done on other established metadata specifications such as the Darwin Core, the Dublin Core, the Federal Geographic Data Committee and DocBook among others. EML extends the functionality of the aforementioned specifications and emphasises the descriptors needed for Biology and Ecology. Particularly, EML enables automated analysis due to its comprehensive data descriptors and the high degree of granularity that facilitates automated content parsing.

An entire book could be devoted to EML, and it is well beyond the scope of this paper. The basics of EML are well covered in Fegraus et al. (2005). Other references (Nottrott et al., 1999; Jones, 2001a; Jones et al., 2001b) provide further detail.

EML is just a container, a portable database that stores metadata records in XML format. How EML is used determines the functionality of the EML record. When a user provides just the minimal information required being compliant with the EML rules (the title, owner and

point of contact), then the functionality of the metadata set remains merely bibliographic. When a user provides more comprehensive and detailed information, then the potential functionality is much richer and machine-mediated data analysis is possible (Ludäscher et al., 2006). In LTER, we call metadata content rich when the metadata includes details of the data location, provenance and data entity structure details such as the physical data storage, the number of headers, delimiters, encoding and the like. Other terms synonymous with content-rich metadata are "data integration level" (Sheldon, 2004) and "attribute level EML". Somewhere in between content-rich metadata and the minimum content required to be EML compliant is a level we call "discovery-level EML". Discovery-level EML (Michener, 2006) allows a person to locate (discover) the associated data. Such content would include a descriptive title, an abstract of the study, some dataset-relevant keywords, a high-level geo-temporal annotation and perhaps a URL to a catalogue containing the data.

Other scientific communities interested in standardising data are developing minimum requirements (MIAME, MIACA, MIMS/MIGS, MIMIX, etc., see Nature's Community Consultation page) for describing data. However, the LTER Network strives to use EML so that the data description is well beyond a set of minimum requirements. LTER promotes the use of EML to create detailed, durable data description records that can be interpreted, integrated and synthesised without much human intervention.

However, the goal of machine-mediated data synthesis, driven by metadata, is quite a lofty vision for the LTER community at this point. In practice, even achieving humbler goals, such as making data accessible and sharing it with others, implies a community commitment and a scientific cultural shift. These human factors have proven to be key factors in the process of LTER metadata standardisation.

4 EML implementations methods

After the official adoption of EML as the vehicle to standardise LTER metadata, each site was left with the responsibility of completing the task. This mandate to transition to EML was unfunded and somewhat controversial among the LTER community. Nevertheless, EML implementation became tied to NSF funding after 2005 when it was added to the sites' review guidelines (Boose, 2005). The responsibility to carry out a site's metadata standardisation (sometimes referred to as EML implementation) fell on the shoulders of the site's information managers. To assist the IMs and coordinate the metadata standardisation project in 2005 the LTER Network Office hired an additional staff member.

In the course of implementing the LTER metadata standardisation process, two factors emerged that defined the implementation strategy per LTER site: technological capability and motivation level. The latter within the LTER community we have referred to as the 'human in the loop' factor.

For purposes of describing the general implementation approaches, we can describe all 26 sites using the following three categories. Classifying sites according to their response to the EML implementation process can have some ramifications (Bowker, 2000). We classify LTER sites in the spirit of community building and we use categories for their power in knowledge building. These categories encompass technical aspects as well as the socio-technical multiperspective. The classifications were built along the process of EML implementation and were not finalised in a formal structure until some time after the LTER reached the project mid-term milestone (San Gil, 2007). Many LTER sites changed their disposition towards the implementation process during the course of the project, often changing their place in this crude categorisation.

- *Early Adopters*. This group included those sites that had both the technological capability to convert their legacy metadata into the EML format and the motivation to complete the task.
- *Building Readiness*. This group actually had the technology capacity to implement standardisation, but for some reason lacked the motivation to complete the task in the short term. This included those sites that considered implementation of a low priority and those sites concerned about the prioritising of site resources.
- *Infrastructure Limited*. This group devoted minimal resources to information technology and while the IM – when one exists – is motivated to complete the task – the requisite resources and expertise is lacking.

Each category described above called for a different strategy to succeed in the metadata standardisation. The early adopters were comprised of a group of about six LTER sites with strong technological resources and enthusiastic information management departments. This group developed sound plans for the standardisation of their metadata. The group did not require (and often declined) any resources offered by the LTER Network. The completion of the site legacy metadata standardisation lasted for about a year, and the sites sent their metadata in the EML format to centralised metadata clearing houses for distribution to the network and the public at large. Some technical details of the process will be revealed later in this section.

The second category, Building Readiness, represented more than a group of standardisation-skeptical LTER sites. It also encompasses a number of sites that were undecided. In a few cases, the site had postponed the decision to implement an EML-based metadata standardisation plan. For the most part, this group of about a dozen LTER sites represented those sites who leaned towards a wait-and-see strategy before investing serious time, personnel and monetary resources to the plan. Concerns raised by these sites included the high cost of the plan, the EML steep learning curve, the questionable benefits (for the site and at

the network level), and concerns about the maturity of EML as well as the maturity level of the software tools accompanying EML. Whatever the reason, the important point is that the site needed more time to decide when and how to implement the standardisation project. A logical metadata standardisation strategy in these sites included providing some incentive to motivate the site to move forward with standardisation. A quick EML discovery-level implementation for a percentage of the site metadata records provided enough motivation for engaging in the implementation process.

Sometimes a multi-pronged strategy motivated a site to embrace EML. Demonstrating EML's value to a site by using application prototypes that were EML-driven was very eye opening to many sites unfamiliar with or skeptical of the potential worth of EML. Another successful strategy was to assist a site in producing a low-cost, discovery-level implementation of their metadata standardisation. Formally recognising the site's efforts in achieving milestones in the process of adopting EML was also a great motivating factor. Another strategy involved periodically providing information of their site's status in relation to the progress of the other sites in the LTER network (San Gil, 2006). This last strategy was particularly effective in motivating the site's principal investigators to take on a more active role.

The third category, Infrastructure Limited, encompassed about six LTER sites. It describes those sites that devoted less than desirable IT resources to their site. This category also contains those sites who, at the time, were not technologically ready to undertake on the complexities of EML, and who believed that the sheer volume of legacy metadata could not be processed by an XML editor or similar entry tools. Whatever the reason motivation was, yet the site information manager did not have the resources to implement. The sites in this category accepted direct help in the form of resources (personnel and technological) from the LTER, via the LTER Network Office. Then, a plan that leveraged the available local resources at the site was implemented. Often, this plan coincided with a transition to EML with discovery-level content, similar to the plans designed for the sites in the 'Building Readiness' category. In this case, the strategy was not to motivate the site, but to maximise the local resources. However, in at least two cases, the site legacy metadata did result in more desirable rich-content EML.

Before we delve further into analysis and lessons learned in the LTER legacy metadata standardisation process, we describe here the different technologies used to accomplish the LTER legacy metadata conversion to EML.

Because of the strong LTER data management policies in place before this project, all the sites had structured metadata for all their studies. Therefore, all technologies used involved developing a customised mapping from a site's existing metadata structure to EML. Fortunately, within a site most of the studies had the same structure, or a small set of very similar structures. This implied that mapping to EML involved developing one mapping per site. A few sites shared very similar metadata structuring templates, and followed similar technical methods to implement EML. Sometimes variations of the same mapping were used, and work done at one site was leveraged by another. That was the case with the Andrews LTER and the Konza Prairie LTER. Synergies also occurred between the metadata structures at Arctic LTER, Plum Island LTER, McMurdo Dry Valleys LTER and Hubbard Brook LTER.

All the sites manage their data and metadata in electronic format, using either a database, or using files in a structured file system. The preferred model for a database is relational. Some sites, however, use a hierarchical database, or a purely XML database or even a hybrid. Many sites relied on a structured file system to organise their data, which was in the form of spreadsheets, word documents, flat text files or even pure HTML in some cases.

In all these cases, the mapping to the EML structure is achievable, and the resulting quality is as good as that stored in the original metadata database or files. The most common method used to transform the file content to EML was Perl scripting in combination with XSLT stylesheets. Sometimes Java or Active Server Pages were used instead of Perl. The first step involved parsing metadata content to XML. Then the XML data was transformed into EML files using a custom XSLT transformation. Some sites developed a custom Excel macro to transform metadata from a spreadsheet to EML. More detailed information about each site's particular metadata management system can be located in the LTER information managers' website.

A good example of the technological aspects of the implementation of the Early Adopters group is the Georgia Coastal Ecosystems LTER site (GCE). This site is managed at the University of Georgia-Athens. The GCE site modified their existing integrated information system to offer EML metadata for every data (Sheldon, 2003). The EML files are dynamically generated EML from the site relational database (SQL server) using the Active Server Pages (ASP) scripting language. EML metadata documents are provided at the two mentioned content levels, discovery level and rich content. At GCE, these levels are called 'Basic EML' and 'Complete EML'. Other early adopters followed a similar technological path – The Andrews forest created active server pages scripts to create EML from their back-end relational database instance (Henshaw et al., 2002). The Central Arizona Phoenix (CAP) LTER site developed custom open-source-based java applications that leveraged XML technologies to exchange information in a semi-automated fashion. Xanthoria included web services using XML-based Service-Oriented Architecture Protocols (SOAP) to take advantage of the EML implementation. Xanthoria used XSLT style sheets to extract metadata from a mySql instance of the CAP relational database. The use of back-end relational databases to manage metadata was a common denominator amongst the early LTER adopters. In general, the implementations differ in the technique used to extract metadata from the database, and make it

EML compliant, although not surprisingly, the processes were always similar.

For sites in the 'Building-Readiness' category, there were just a few differences on technological aspects of the EML. For example, the low-cost, discovery-level strategy was often implemented by rapidly developing a Perl parsing script on the most common structure metadata format. The Sevilleta LTER, the Arctic LTER and Plum Island Ecosystem LTER had structured metadata in a text-based document format. A Perl script parsed the basic metadata and transformed it into basic XML records. Then, an XSLT reformatted the XML metadata to make it EML-2.0.1 schema compliant. The Luquillo Experimental Forest LTER and the Short Grass Steppe LTER had a simple database where discovery-level metadata was stored. For these sites, the conversion involved exporting the metadata into raw XML tables that were parsed with a combination of a Perl script and a XSLT style sheet. The Perl code managed the relationships amongst the pieces of the metadata records. The Perl code created a simple XML file that contained all information available for a single dataset. We applied an XSLT transformation to the simple XML file in the next step of the implementation workflow. This XSLT transformation resulted in an EML-compliant XML document, and this process was repeated in an automatic batch process.

In both cases – basic metadata in text-based format and databases – some post-transformation quality checks followed to ensure the quality of the resulting EML records. Minor errors were corrected to ensure that all metadata records were EML schema compliant. In addition, we revised manually a random subsample of about 10% of the total resulting records to verify the content. For these sites, we planned a second step in the process of implementing content-rich EML. This second step involved modifying slightly the existing Perl scripts and style sheets to accommodate rich-content metadata. Frequently the process had to be complemented with some restructuring of the legacy content-rich metadata. Thus both the Sevilleta LTER and the Liquidly LTER sites a bit over a calendar year correcting and reformatting existing metadata to make it compatible with a structured source template. An EML transformation of the template was created using Perl and XSLT as mentioned. The metadata correcting process included revising metadata, providing missing information: from missing abstract, to definitions of variables, codes used to identify plots and species and scientific units. We also checked the consistency of the data file column label headers and the corresponding definitions and labels provided in the metadata record. Finally, we reformatted the structured files for compliance with the common template. Luquillo LTER corrected over one 100 metadata sets manually, over the course of a two-year period.

The building readiness McMurdo Dry Valleys LTER site, the Coweeta LTER site and the Cedar Creek Natural History Area LTER sites adopted a more direct process to enrich their metadata from their legacy format. These three mentioned LTER sites had structured metadata that included raw text files and HTML files. Like in previous cases, we created custom Perl parsers to prepare EML-compliant rich-content metadata, and produced EML after creating a simple XML as intermediate step, except in the Cedar Creek site, where a direct transformation from the legacy metadata to EML was applied using Perl custom code.

All the Perl scripts and styles sheets are available through the LTER Network Office at http://urban.lternet.edu/viewvc/trunk/conversions/?root=NIS.

We found out that it is relatively easy to convert or transform structured metadata. Altova's MapForce graphic tool helped us mapping a structured metadata to EML. This graphic tool allowed us to automatically generate scripts (XSLT or Java) to transform content into the EML schema. MapForce has a drag-and-connect feature to associate information placeholders from different structured containers. Correcting errors that are traced to the instance of metadata documents that do not conform to the common mapped structure is time consuming, albeit easy. The EML implementation process had however more involved issues. Collecting missing legacy metadata proved time consuming and sometimes futile. Some datasets over ten years old were lacking critical information that was hard to retrieve. For example, at Coweeta LTER, a research site that started watershed experiments in 1934, we spent about three calendar months contacting principal investigators and associate researchers to collect critical missing metadata for some of the 200 research projects listed at the Coweeta data catalogue. This process involved reading related publications, and producing educated questions for the project associates. We reviewed all the projects, at a rate of about ten projects per day. We received many positive responses. About half of the contacted people solved the missing or incorrect metadata information within a week or less, while about 20% solved the metadata problems partially and vowed to collaborate. Only about less than a quarter of the projects contacted failed to produce a satisfactory resolution. In one case, the principal investigator passed away just a week before being contacted about his research.

5 Discussion and lessons learned

In this section, we discuss further the strategies adopted to implement the LTER EML adoption plan and attempt to convey some lessons learned in the process. We also discuss the advantages and disadvantages of utilising a particular technology in relation to the site resources and the level of motivation.

We broadened the horizon of potential implementation strategies by including social, "human in the loop" factors along with technological factors, instead of merely seeking the most efficient technical strategy. To gain a site's full support for the EML implementation project, we provided convincing evidence of EML's effectiveness to the site information manager and principal investigators. We gained the trust and motivation needed to complete the project

by using examples of data synthesis to substantiate the EML advantages advocated by the LTER leadership.

Once the "human in the loop" factor was identified as critical to project success, a more complete socio-technological framework guided the strategies used to carry out the standardisation project. We designed a strategy that focused primarily on addressing the weakest factor at each site. A strategy that motivated a site to action or provided technical resources eased initial resistance to the project and then cleared the path to completion. It is important to note that we often dismissed more efficient technical implementation plans to address primary concerns. The shortest path, we learned, was not always the best path.

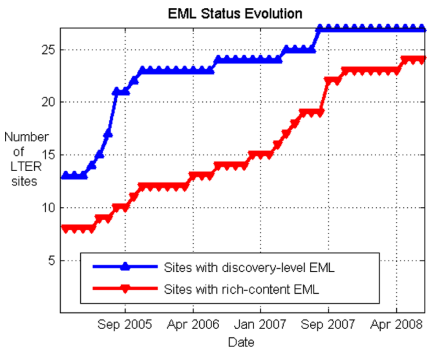
We categorised the sites according to their 'readiness' with regard to level of technological resources and degree of motivation. That is, how ready was a site to implement, and what factor – whether technological or social (Karasti and Baker, 2004) – was holding them back from completing the task of converting the legacy metadata to EML. One could argue that these categories are sometimes related by causality, and are thus not completely independent. Is there a lack of motivation due to sparse technological resources? Are technological resources thin due to a relative lack of interest in information management? It is hard to answer, and some sites threaded both categories. Within these categories one could attempt to rank the degree of motivation, IT resources and so on, creating a highly discrete category domain, but any result would be highly subjective. The usefulness of the categories was that they explain the need being flexible in the way implementation is carried out. That is, one strategy did not fit all sites. In addition, in many cases the strategy adopted was similar whether the site lacked motivation or lacked resources. In either case, these sites began with relatively easier to achieve goals, such as creating EML at the discovery level for a fraction of their metadata, before moving the bulk of their data to rich-content EML (San Gil, 2008).

This paper focuses on the LTER site's EML implementation, but it is worthwhile to mention the prior work done by the ecological community that laid the foundation for EML adoption by the LTER. The adoption process, which is documented in the LTER committees meeting minutes (found in the document archives – at the intranet web page for lternet.edu) would also be valuable information for communities and organisations in the process of adopting a community metadata standard. In our case, the lesson learned is that even though EML was deemed the best solution for a network wide standard, there were certain risks inherent in its selection. EML has a steep learning curve. Thus monetary and personnel resources had to be allocated to actually implement the project. The sites initially regarded the benefits of EML adoption as long term and largely falling outside the site, while consuming some local resources. There were some concerns about the maturity of EML, as well as the maturity level of the software tools accompanying EML. Whatever the reason, the important point is that the lack of motivation threatened any implementation strategy, and it was critical to identify these human readiness factors before launching an implementation plan.

One important factor that energised the LTER EML implementation was reaching the critical mass milestone. At the 2006 LTER Information Managers meeting, the authors presented the "EML implementation status of sites" (See Figure 1). At the same time, half of the LTER sites had rich-content EML while about 90% of the LTER sites had some metadata records in EML form with discovery-level content. Graphical views in the form of Venn diagrams of the project status of the LTER sites, as well as tabular quantitative metrics of the standardisation project progress, conveyed the relative position of each site in the project progress. Succinct peer pressure, along with some strong emphasis on each site's achievements, always resulted in a spike of information managers' time devoted to the EML implementation project. At the same meeting, some prototypes of EML-driven data analysis were presented, which helped some reluctant sites to push their EML content beyond the discovery-level requirement recognised by NSF.

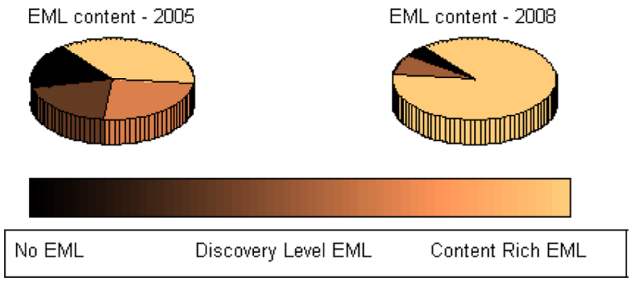
On Figure 1, the fast growing top curve (upright triangles) shows the pace at which LTER sites implemented EML at least at a discovery level; that is, providing enough information to make the data searchable and discoverable. We can interpret the top curve (upright triangles) as the symbolic adoption of EML by the site. Within two years, the same sites that accomplished some degree of EML implementation earlier (top curve (upright triangles)), motivated themselves and provided rich-content metadata in the EML format, which is shown by the bottom curve (inverted triangles). To date, over 90% of the LTER sites have rich-content EML, and some of the sites have subjected their metadata to strict quality control procedures.

Figure 1 Details the number of sites implementing EML over time. In this figure we can see the combined effects of motivating sites and reaching critical mass (see online version for colours)



The pie charts depicted in Figure 2 show the overall state of LTER metadata. In dark colours we see an estimate of the metadata that has not yet been transformed to EML, and as the colours brighten, we see the estimated relative proportion of content-rich metadata implemented in EML. The beige shades in between represent the metadata in EML with a discovery-level content or less.

Figure 2 Estimated proportions of LTER metadata converted into EML records of different content richness. The two charts correspond to summer 2005 and summer 2008 respectively. Darker colours indicate no standardisation or mere bibliographic record quality. Brighter colours indicate higher level of content in the record, up to the defined 'content-rich' level (see online version for colours)



We realise the importance of evaluating this effort in a cost-benefit, value-justifying framework (Lytras and Sicilia, 2007). Thus, we provide cost estimates based on the time invested in the EML implementation. However, the cost estimate must be limited to the metadata conversion project, and not to the additional costs of metadata capture, the legacy metadata management systems, and the decision-making process to adopt a specific language (EML). The information managers (co-authors of this paper) have had extensive discussions about the value of providing limited cost estimates here. We should emphasise the perils of extrapolating the estimates below, as there are many components that we have not attempted to quantify. Can we still justify the high cost of the metadata-related expenses? Perhaps our best argument in favour of them is to view quality, standardised metadata as a critical network necessary to function as an integrated entity. The authors conclude that while it is necessary to asses the overall cost-benefit of the metadata project, it is well beyond the scope of this paper. We cannot estimate the overall metadata cost since it involves many decades and many changes of information management systems and personnel.

Here, we provide cost estimates for the EML actual implementation, that is, the cost of providing EML from the legacy metadata systems. On average, each LTER site in the building readiness category invested about a month of the data manager's time and a month of the project coordinator's time to provide a basic conversion with basic quality control mechanisms such as EML schema compliance and some deeper random quality checks. Further quality checks cost more time. We mentioned up to two years of calendar time, or about three dedicated months of the local site data manager and the project coordinator's time. Notable deviations happened, as mentioned before, some sites were more ready than others, even in this category. Some data managers took longer (up to two months per basic implementation), and some were faster (two weeks). The project coordinator also gained experience as time progressed, which helped to speed up the implementation process.

Some of the EML early adopters LTER sites took about the same time. However, these sites did not benefit from the expertise of the project coordinator. The Konza Prairie LTER information manager reports that the basic conversion took about two months of the data manager's time to attain discovery-level EML and another month to enhance to EML compliant, content-rich EML. However, the data manager reports about a year and half to make metadata corrections necessary to provide machine-mediated functional EML. The Georgia Coastal reports an EML implementation process of the order of eight days. The Georgia Coastal Ecosystems LTER Information System includes a database-driven management system with data-derived metadata and quality checks that was in place before the EML implementation project – a huge advantage that came with an unspecified cost. Another early adopter site with ample expertise is the Virginia Coastal Reserve. The Virginia Coastal reports a solid month of work spread over several months to provide fully compliant, rich-content EML.

In Summary, many sites have reported estimates in-line with the crude time estimates provided here. However, those estimates are limited to the aspects of refactoring existing metadata into EML, and as such should be considered by the reader.

Another important lesson learned is that when building a community-based IT project, it helps to recognise and address the human factors. In fact, the technical aspects of the project fell into place after the human aspects had been addressed.

6 The road ahead

While the project reached an important milestone – all LTER sites are contributing with most metadata in EML format – it is not yet complete. The final goal is to have all LTER metadata standardised in the EML format. More specifically, we strive to achieve rich-content EML for all the metadata whenever possible. A handful of LTER sites are yet to finalise the EML migration project as of May 2008. Quality control aspects of the metadata have not been discussed in this paper. But the authors are aware of problems with some of the EML metadata documents, mostly due to a lack of an in-depth quality control process. At the time of submission, at least three sites were engaged in a full revision on the accuracy of the metadata. Some of the problems detected in the metadata arise from the indirect nature of the data documentation. Many of the EML metadata records are described a posteriori, as opposed to being derived directly from the data by some automated processing (such as machine parsing).

The authors submitted a manuscript highlighting the most common uses of EML in LTER or EML in practice (San Gil and Vanderbilt, 2009). These common practices complement and sometimes are in contrast with the LTER Best Practices and the EML usage guidelines. We will discuss rarely used or non-used information placeholders in EML, and address practices that deviate from the EML official guidelines. Also, we highlight metadata needs that are not met by EML. For example, all genomic-related

metadata are not adequately described in EML, and with the advent of eco-genomics the need of a proper metadata standard is critical. We will also discuss the process of enhancing EML, and the role of the community in such a process.

Acknowledgement

Inigo San Gil would like to acknowledge the support of the National Biological Information Infrastructure through the Cooperative Agreement with the Long-Term Ecological Network. Inigo San Gil is also grateful to Alicia San Gil for providing deep insights on the focus and format of the manuscript.

References

Anderson, C. (2008) 'The end of theory: the data deluge makes the scientific method obsolete', *Wired*, Vol. 16, pp.116–121, http://www.wired.com/science/discoveries/magazine/16-07/pb_theo

Boose, E. (2005) *LTER Site Review Guidelines – Information Management Review Criteria*, Information Managers Executive Committee, http://intranet.lternet.edu/im/im_requirements/im_review_criteria

Bowker, G.C. (2000) *Sorting Things Out: Classifications and Its Consequences*, MIT Press.

Fegraus, E.H., Andelman, S., Jones, M.B. and 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*, Vol. 86, pp.158–168.

Harmon, M. (2003) *Motion to Adopt EML at the Coordinating Committee*, http://intranet.lternet.edu/archives/documents/reports/Minutes/lter_cc/Spring2003CCmtng/Spring_03_CC.htm

Henshaw, D.L., Spycher, G. and Remillard, S.M. (2002) 'Transition from a legacy databank to an integrated ecological information system', *The 6th World Multiconference on Systemics, Cybernetics and Informatics*, International Institute of Informatics and Systemics, Orlando, FL, pp.373–378.

Jones, M.B. (2001b) EML Specification, http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html

Jones, M.B., Berkley, C., Bojilova, J. and Schildhauer, M. (2001a) 'Managing scientific metadata', *IEEE Internet Computing,* Vol. 5, No. 5, pp.59–68.

Karasti, H. and Baker, K. (2004) 'Infrastructuring for the long-term: ecological information management', *Hawaii International Conference for System Science Proceedings. of the 37th Hawaii International Conference on System Sciences, HICSS38*, IEEE Computer Society, January, Hawaii, USA, pp.321–358.

Lternet website (2008) Abridged Lternet History at http://lternet.edu/about/history.html

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J. and Zhao, Y. (2006) 'Scientific workflow management and the Kepler system', *Concurrency and Computation: Practice and Experience*, Vol. 18, No. 10, pp.1039–1065.

Lytras, M. and Sicilia, M. (2007) 'Where is the value of metadata', *International Journal of Metadata, Semantics and Ontologies*, Vol. 2, No. 4, pp.235–41.

McCartney, P. and Jones, M.B. (2002) 'Using XML-encoded metadata as a basis for advanced information systems for ecological research', *Proc. 6th World Multiconference Systemics, Cybernetics and Informatics, Vol. 7, Int'l. Inst. Informatics and Systematics*, Orlando, FL. USA, pp.379–384.

Michener, W.K. (2006) 'Meta-information concepts for ecological data management', *Ecological Informatics*, Vol. 1, No. 1, pp.3–5.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B. and Stafford, S.G. (1997) 'Nongeospatial metadata for the ecological sciences', *Ecological Applications*, Vol. 7, No. 1, pp.330–342.

Nottrott, R., Jones, M.B. and 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.

San Gil, I. (2006) *EML Status of the LTER Sites, Data Synthesis Examples and the Next Step*, Databits Fall 2006 Issue, http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06fall/#7fa

San Gil, I. (2007) 'LTER to meet metadata milestone this summer', *LTER News 2007*, Vol. 2., http://www.lternet.edu/news/Article150.html

San Gil, I. (2008) *LTER Metadata Information System Profiles*, http://intranet.lternet.edu/im/siteprofiles

San Gil, I. and Vanderbilt, K. (2009) 'Two examples of ecological data synthesis driven by quality EML, and the practical guides to use EML to this end', *Special Issue on Ecological Informatics, In review*.

San Gil, I., Sheldon, W., Schmidt, T., Servilla, M., Aguilar, R., Gries, C., Gray, T., Field, D., Cole, J., Yun Pan, J., Palanisamy, G., Henshaw, D., O'Brien, M., Kinkel, L., McMahon, K., Kottmann, R., Amaral-Zettler, L., Hobbie, J., Goldstein, P., Guralnick, R.P., Brunt, J. and Michener, W.K. (2008) 'Defining linkages between the GSC and NSF's LTER program: how the Ecological Metadata Language (EML) relates to GCDML and other outcomes', *OMICS: A Journal of Integrative Biology*, Vol. 12, No. 2, pp.151–156.

Sheldon, W. (2003) *The Georgia Coastal Ecosystem LTER Information System*, http://gce-lter.marsci.uga.edu/public/data/eml_metadata.htm

Sheldon, W. (2004) *EML Best Practices*, http://cvs.lternet.edu/cgi-bin/viewcvs.cgi/*checkout*/emlbestpractices/emlbestpractices-1.0/emlbestpractices_oct2004.doc?rev=1.1

Websites

LTER History, http://www.lternet.edu/about/history.html

National Biological Information Infrastructure, http://nbii.gov

EML-Developers, http://knb.ecoinformatics.org/software/eml/members.html