Examples of ecological data synthesis driven by rich metadata, and practical guidelines to use the Ecological Metadata Language specification to this end

Inigo San Gil*

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

Kristin Vanderbilt

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

Steve A. Harrington

SAHarrington 70,
West St. Middletown,
Spriens VT 05757, USA
E-mail: scribes1963@yahoo.com

Abstract: We present two examples of scientific results using a semi-automated data synthesis driven by quality, rich-content metadata: 1) Antarctic climate and 2) effects of drought on biodiversity. We use a framework for semi-automated processing of data supported by quality controlled, content-rich metadata expressed in the Ecological Metadata Language (EML). We discuss a set of common practices for EML newcomers as a valuable guide for the EML use. We provide some simple tools that can be used to address quality control as the EML is generated. Based on our extended EML experience, we make recommendations about the future of EML.

Keywords: metadata; EML; ecological metadata language; LTER; long term ecological research network; climate; drought; biodiversity; data synthesis; data integration; machine mediated data analysis.

Reference to this paper should be made as follows: San Gil, I., Vanderbilt, K. and Harrington, S.A. (2011) ‘Examples of ecological data synthesis driven by rich metadata, and practical guidelines to use the Ecological Metadata Language specification to this end’, *Int. J. Metadata, Semantics and Ontologies*, Vol. 6, No. 1, pp.46–55.

Biographical notes: Inigo San Gil is the liaison agent between the National Biological Information Infrastructure and the LTER. He coordinates metadata interoperability among the NBII and LTER programmes. He leads a team of ecological information management that develops solutions based on the Drupal content management system.

Kristin Vanderbilt is an Associate Research Professor in the Department of Biology at the University of New Mexico. Her interests include ecoinformatics and statistics. Since 2000, she has served as the Information Manager for the Sevilleta Long Term Ecological Research (LTER) programme, a desert research site in the Southwestern United States.

Steve A. Harrington is a graduate of the Yale School of Forestry and Environmental Studies, former National Coordinator of the Forest Guild, and co-author of the Four Corners Assessment on forest fires and fuels in the Southwestern United States.
1 Introduction

Metadata is an increasingly important tool in ecological research, with utilities that are advancing beyond the traditional purpose of data preservation (Lytras and Sicilia, 2007; Michener, 2005). The LTER actively applies structured, standardised, and portable metadata technologies for data analysis, integration and sharing, as well as digital interoperability.

In this paper, we illustrate how high quality metadata can contribute to science with two examples that recreate published scientific results using a metadata-driven prototype. The first uses a study of an Antarctic local cooling effect, published by Doran et al. (2002) that revealed a surprising climate cooling trend during the last two decades of the last millennium. We used a simple process to replicate and expand upon the published results. The remarkable part of this analysis is that it took very little time and required minimal communication with the original authors. Success in an extended analysis process is dependent upon accessible, high-quality data and metadata.

In another example, we recreated and extended the results of a published analysis (Tilman and El Haddi, 1992) on the effects of drought on biodiversity. These two examples served to motivate parts of the ecological scientific community to the task of creating quality metadata.

This paper is divided in two parts. The first part is about the application of EML documents. The second part is about the practices on the creation of EML documents. We learn about how to make EML so it can be used in metadata-driven studies such as the cases shown in this paper.

In this paper we will explain the details of this prototype and the aforementioned analysis as well as the underlying metadata specification used. We will also outline some of the common practices for using one particular metadata standard, called EML. Finally, we analyse the quality to date of the of the LTER’s EML.

The mission of LTER (2011) 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. LTER unites over 2000 scientists conducting research on 26 sites located across the continental USA, Alaska, the Caribbean, the Pacific and Antarctica. There are over 40 other nations that have formed LTER networks under the umbrella of a global network called International LTER (ILTER). The reader should note that we use the term ‘LTER’ throughout the paper to refer to the US-LTER network.

The EML is a scientific metadata container implemented using the extensible markup language (XML) schema. The basics of EML are detailed in Fegraus et al. (2005). McCartney and Jones (2002) and references therein offer further insight on EML as the facilitating tool to build a federated network of data. In 2003 the LTER adopted (Harmon, 2003) the EML as the standard for LTER network metadata. All sites are providing EML documents to the central data catalogue, but the conversion of LTER metadata to EML is not yet completed (San Gil et al., 2009).

EML is essentially a container, a portable database that stores metadata records in XML format. EML is also a vehicle for interoperability between different ecological data providers and consumers. It is the quality of the metadata itself, expressed in EML, that dictates how useful the EML document can be. Providing EML content that meets the minimum information (the title, owner and point of contact) required for compliance with EML rules merely enables a bibliographic use of the EML document. It provides a record for the data, but requires further investigation to actually acquire the data. Machine-mediated data analysis functionality can only be achieved by providing more exhaustive content (Ludäscher et al., 2006).

Similar functional-oriented categories for metadata are also defined and explained by Michener (2005). We define metadata capable of machine-mediated data analysis as rich-content EML. That is, a EML document is rich-content when the associated data can be consumed (parsed and processed) by an automated procedure, such as a web service client, an automated workflow or other programme of this nature). Rich-content EML includes details for the data container structure, such as: the physical data storage, number of headers, physical location (URL), units and codes used, as well as other basic contextual information about the data. We assume that rich-content EML is also high quality, meaning that the metadata has undergone a thorough review to purge errors. We define metadata that falls somewhere in between rich-content metadata and the minimum content required to be EML-compliant as discovery level EML. Discovery level EML allows a person to find (hence discover) the associated data described by the EML document. 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. All LTER sites have attained discovery level EML for almost all their datasets, and many sites offer rich content EML. The LTER sites are required to provide at a minimum discovery-level EML, although it is also encouraged to provide rich-content EML.

2 Examples of metadata-driven data analysis

We illustrate the value of structured and standardised quality metadata using two examples. These two particular cases were chosen for two reasons: They are hot topics of research (global warming and biodiversity), and because of the availability of rich-content metadata in EML format associated with the data. Also, as the underlying subject matter is of general interest, the examples are understandable by scientists trained in disciplines other than the purely ecological sciences.
2.1 Climate cooling revisited

In an Antarctic climate study published in *Nature* in 2002 (Doran et al., 2002), Peter Doran showed how the average air temperature at the Lake Hoare weather station in East Antarctica dropped over the last 20 years of the last millennium at a rate close to the estimated rate of global warming. Given prevailing global warming trends, we were surprised to encounter such findings. We set out to replicate the analysis, and then extended the analysis five more years to include more recent data readings.

Gathering the data the traditional way would have taken some time, and would likely have involved trying to contact the principal investigators, and then asking them a number of detailed questions. With rich-content EML now adopted and implemented in the site, we were able to locate the data by performing a search on ‘Lake Hoare’, using the LTER data catalogue web interface, which resulted in two records (see Figure 1).

Figure 1 Screen showing the records returned by a simple search on ‘Lake Hoare’ in the LTER data catalogue

<table>
<thead>
<tr>
<th>Title</th>
<th>Contacts</th>
<th>Organization</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake Hoare Air Temperature Measurements</td>
<td></td>
<td></td>
<td>temperature, air temperature</td>
</tr>
<tr>
<td>DORLTER-001-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lake Hoare Biologic Site Measurements</td>
<td></td>
<td></td>
<td>biology, biogeochemical</td>
</tr>
<tr>
<td>DORLTER-001-4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The LTER data catalogue website allows the public to access the metadata and interpret the associated data format and context. We used a simple customised Perl parser programme to retrieve the associated Lake Hoare air temperature data automatically, and format it for analysis with *Matlab*. Once the relevant data was loaded in *Matlab*, we plotted the data mimicking the graph published in the *Nature* paper (see Figure 2). We soon realised that the values did not quite match what was published for the overlapping range of years (1994–2000). At that point, we contacted the appropriate data manager for a simple inquiry that was resolved in a day. The information needed was the seasonal time-frame definition for the McMurdo Dry Valleys LTER in Antarctica. The journal article (Doran et al., 2002) did not include sufficient details about the data to replicate the analysis. The basic methodology was to take in the 15-minute temperature measurements, and calculate the average for all four seasons (Summer: December January February; Autumn: March April May; Winter: June July August and Spring: September October November). In the original study and this replicate study, we take each seasons average and subtract the overall season average from it (i.e., how much does the Spring average in 2004 deviate from the average of all Spring seasons for all time). Finally, we plot all the seasonal deviations – there are 4 points per year. Once the averages were computed over the same ranges, the recomputed graph matched the original published graph for all 24 points but one. The extended range did not significantly modify the new least-square fitting, and actually continued the trend observed by the original authors.

Figure 2 Three graphs of the surface air temperature at McMurdo Dry Valley’s LTER. The top graph is a replica of the plot shown in Doran et al. (2002). The new analysis shows in the middle graph an extended plot range, showing a good match with original. The bottom graph shows aggregated air temperature data from nearby stations

To further validate the results of the Doran study (and to further illustrate the value of EML), we retrieved data from nearby weather stations within the McMurdo Dry Valleys LTER in Antarctica. The additional seven air temperature records can be considered samples of the same temperature record for the purpose of evaluating the temperature trends over time. The analysis provided about the same rate of cooling for the period 1994–2005. The aggregated temperature time series can be seen in the bottom graph of Figure 2.

Note that the choice of technologies to replicate the analysis could have been different; these basic statistics could have been computed using *R*, *SAS*, *S-PLUS* or *Excel*, the point of the exercise was to show the value of metadata through a case example of data synthesis. Furthermore, the analytical process can be integrated in a scientific workflow software of choice.

Thus, we were able to efficiently re-analyse previously published results and extend those results with more recent data from the same, continuous data stream. We validated the original results and further corroborated them with new information (temperature records) from the same weather station and nearby ones. The availability of quality metadata compliant with a well-established specification, the EML, facilitated the data integration and analysis process, and also supported the aggregation of data to the same overall study.
from other weather data streams that are documented in EML and stored in the LTER data catalogue or in some of its mirroring servers. Eventually, the general public will be able to use these prototyped tools through the EcoTrends portal (Brunt, 2006), which is slated to go public sometime in the near future. The EcoTrends portal is a portal into the long-term data series from LTER, the experimental forests research sites from the US Forest Service and others. The corresponding author of this paper proposed the idea of a dynamic web-portal to the Trends Editorial Committee, an ad-hoc committee created to edit and publish a book on long-term datasets. Even though the idea was not initially warmly received, the author was able to pursue the project with the help of a technical committee led by Dr. Servilla. The author developed several web-portals for gene expression time-series (Arbeitman et al., 2002; Reinke et al., 2004). Initial prototypes were presented at the LTER All Scientists meeting at Estes Park, and the EcoTrends web-portal was completed on 2008.

2.2 Biodiversity in grasslands

For our second example we used a simple study by Tilman and El Haddi (1992) that demonstrated the negative effects of drought on species richness in a plot located in the Cedar Creek Natural Reserve LTER site, about 35 miles north of the Minneapolis – St. Paul metro-area (see aerial photo in Figure 3).

Figure 3 Aerial photo of the experimental plots at the Cedar Creek Natural History Area LTER. A series of long term research studies have been conducted in this area

We searched for the terms “long-term aboveground biomass stability diversity productivity” in the LTER catalogue’s advanced search, and obtained 25 datasets corresponding to the year-by-year datasets associated with this study-plot. For this study we again used a Perl code to download and re-format the data. The data can also be found in the LTER data catalogue (see Figure 4).

In this case, the data are spread over 25 different data files corresponding to the 25 years of the experiment that began in 1982. By interpreting the metadata, we were able to integrate all the data into a single file ready for analysis in Matlab. Once in Matlab, we generated two time series graphs that mimicked the graphs presented in Tilman and El Haddi’s (1992) publication. Both the cumulative precipitation over time and the species richness over time showed perfect agreement with the published results (see Figure 5).

The graphs show how biodiversity, measured here as species richness, was negatively affected by two drought years. In our expanded plot range, we show how species richness seems to recover close to original levels in subsequent wet years, but later on total species richness seems to be stable at a value ten per cent less than the value shown in the early, pre-drought years. Both fertilised and unfertilised plots show similar correlations between drought and biodiversity.

Reanalysing Tilman’s biodiversity work was straightforward because of the accessibility of standardised metadata in EML. Extending the original analysis was just as easy. Similar re-analysis can be done in the same fashion, which facilitates the scientific review process as well as the advance of science through simple, semi-automated data sharing. Frequently, when data managers and scientists are shown an example of metadata-driven analysis, as described above, they react by asking, “How do we start implementing EML?”.

The following sections respond to that question by providing a framework to develop rich-content EML.
Figure 5  The top graph shows the yearly precipitation at Cedar Creek LTER. The bottom graph shows the number of species over time. The vertical bar at 1990 marks the end of the original work (Tilman and El Haddi, 1992) and the extended range added here.

3 Practical steps in creating EML-metadata documents usable for metadata driven applications

In this section, we describe specific steps to develop rich-content EML documents within the context of the LTER EML implementation process. Note, we do not address the pros and cons of the existing metadata editors, except to note the range of decent ecological metadata editors on the market, from Morpho (Higgins et al., 2002) to Metavist (Rugge, 2005), or XML-based editors such as oXygen (The Oxygen XML Editor, 2011) or XMLSpy (The XMLSpy XM Editor, 2011). Our primary focus is the content ultimately provided in EML.

3.1 Follow the EML guidelines and read the LTER EML best practices for further advice

The EML specifications (The EML Guidelines, 2001) provide definitions and examples of EML content for most of the EML information placeholders. An information placeholder refers to an indivisible unit of metadata content, describing a piece of information that cannot be broken down into further units. Some examples of information placeholders are a zip code, a unit of measurement, or a keyword. Other terms often used interchangeably with information placeholder are XML element or tag.

EML is quite general and broad in scope. Often EML is branded as a scientific language, not just an EML. This non-case-specific nature of EML is desirable for cross-synthesis studies. At the same time, the loss of specificity generates ambiguities that undermine the utility of certain placeholders, which have to be resolved by the user. For example, Julian dates can turn up as nominal measurements, integers, or simply dates.

The LTER users created an EML Best Practices document (The EML Best Practices Document, 2004) to address some of the ambiguities detected early in the LTER EML implementation process. The Best Practices document incorporated some examples of EML use and introduced tier-levels for the functionality of the metadata content. The Best Practices seek to maximise interoperability of datasets and minimise heterogeneity of the metadata content interpretations. Each best practice includes a concrete example and provides an explanation of the rationale behind the practice. The examples serve two functions: practical guidance for those unfamiliar with EML and removal of ambiguities.

One example is guidance on how to use the EML packageID. The packageID, as defined by the EML Guidelines, is a unique identifier for each EML document. The EML guidelines do not provide further information about the packageID. The LTER Best Practices recommends a homogeneous structure in order to provide for easy branding and provenance of EML documents. The recommended structure is composed of a scope, a unique numeric identifier and a version of the EML document. For example, the metadata referenced in the climate cooling study is knb-lter-mcm.7011.6 (see Figure 1), where knb-lter codes all LTER sites, mcm codes the McMurdo Dry Valleys site, 7011 is the unique numerical identifier for the dataset and 6 is the version of the EML record.

There are more recommendations in the Best Practices document, but even following those still leaves a number of ambiguities. Here we examine some troubles with EML and the manner in which some LTER sites have resolved them.

3.2 Practical implementation of EML, learning from the LTER experience

EML offers over 400 information placeholders that can be repeated as the user sees fit. Potentially, a single EML document can describe umbrella projects that encompass several studies and the many data containers that reflect the actual studies. EML’s broad scope and flexibility translates into extensive metadata capture forms that may seem overwhelming to a person attempting to create metadata. EML does not define what a dataset is – it leaves the dataset definition and scope to the EML user, resulting in some inconsistent EML. For example, some users define a dataset as the data stream captured by a single sensor during the span of a year (see EML documents from the Arctic LTER site), while other sites (i.e., the Santa Barbara Coastal LTER) interpret a dataset as a project encompassing multiple experiments, and numerous data streams spanning several years. LTER’s implementation of EML generated a wealth of experience for enhancing EML functionality.

3.2.1 Boundaries to narrow the scope of a dataset

The lack of a unified, coherent vision of the appropriate scope of a dataset poses interpretational challenges for
data-seekers. We refer to this open-ended problem as the ‘lumpers vs. splitters problem’, where the lumpers are the providers that condense multiple projects into a single EML document, and the splitters are those who create a document for each single variation of the same data collection procedure. In the best case scenario, one would first determine the scope of the dataset and then create the content of the metadata document accordingly. The most common EML document at LTER describes a single project or dataset. Each dataset has more than one data container or data entity associated with it. The authors believe that creating multiple EML documents for multiple, different studies enables clear differentiation between related but different studies. EML allows the aggregation (lumping) of multiple studies together in one EML document, however, the unintended outcome of the study aggregation may obfuscate the study discovery. On the other end of the dataset spectrum is the division of studies into their atomic parts. The 25-year-long study on drought and biodiversity was presented in 25 different EML documents. Our query on ‘long term plant diversity’ resulted in 25 hits. The differences between the 25 resulting records were minimal, mainly relating to the time range of the study. In contrast, for our Antarctic example, only one document was retrieved for our ‘Lake Hoare’ query. The resulting record described the research spanning all time ranges available. EML’s flexible characterisation of a dataset leaves the metadata provider with ample, but at times excessive, room for interpretation. The authors realise that in part, the particular discovery problems described are also related to the architecture of the metadata catalogue, not so much with the vehicle of metadata transport, EML.

At LTER, we have metadata providers that work both ends of the dataset-scope spectrum described here. We propose narrowing the scope of datasets to the middle ground interpretation between the described extremes – neither big umbrella projects nor the basic atomic units of a study. The McMurdo Dry Valleys weather station’s LTER exemplifies a possible middle ground scenario. On the one hand, there are many weather stations that acquire similar data, but the associated metadata records are maintained as separate, unique identities. However, though each weather station employs an array of measurement instruments, resulting datasets are not stored on an instrument by instrument basis in numerous EML records, but are instead kept in one unified record.

3.2.2 The problem with classifying a measurement

For the purpose of categorising measurement, the creators of EML followed a four-fold 1946 classification by Stevens (1946): nominal, ordinal, interval and ratio. A fifth category was later added, called the date-time measurement category. However, LTER found the Stevens classification system subject to a range of ambiguities since there is no clear boundary between the classification types (Velleman and Wilkinson, 1993). We found many instances where the metadata provider classified similar measurables into different categories. Some of these measurement classification instances may be attributed to errors, but the volume of the misclassified measurables may be exacerbated by the inherent ambiguity in Stevens’s categories. For example, the ‘Plot number’. an identifier used to differentiate plots where a study is being conducted, has been characterised variously as a nominal, ordinal, interval and even ratio category.1

To reduce the occurrence of such disparities, many within the LTER community have reduced Stevens original four elements (nominal, ordinal, interval and ratio) to two types (nominal and ratio). LTER uses the ratio as the umbrella category for numbers, more specifically numbers that represent ratios of quantities or any quantifiable measurement (e.g., temperature, salinity, censuses). The nominal category is used for every other variable or parameter type that is not a numerical measurement or a date-time measurement. The nominal category includes strings, text, codes, and even numbers where they do not determine a quantity for an actual measurement, such as a plot number. Conversely, the ratio category would include any data categories to which an algebraic operation can be applied. Within the LTER community, many agreed to use a date-time measurement category to classify any parameter or variable that represents a calendar date. Thus, the common practice used by LTER is the classification of measurement types into three categories: nominal, ratio and date-time. This simplified LTER approach to measurement types is not the solution to the overarching problem of classifying measurements, but it greatly clarifies the data entry process without compromising software applications.

3.2.3 EML elements barely used or never used in practice

It may be helpful for new EML users to understand the relative utility of EML elements. The LTER data catalogue hosts about 6000 EML records from the 26 LTER sites, along with other EML documents from other sources. We used a simple Perl script to detect the presence of specific elements from the EML schema in the LTER data catalogue, in order to develop a list (presented below) of those EML elements rarely used or unused, and therefore deemed of low utility.

EML is based on XML, which can be viewed as a portable database organised into an hierarchical structure. The top of the EML catalogue offers descriptors for data sets, software packages, bibliographic citations and scientific research protocols. Within these four categories, EML offers a number of relevant elements, some of which appear within multiple categories. We quantified usage of the four top-mentioned categories and discovered that most of the LTER EML documents only use the dataset EML module. In contrast, the EML software module is not used in any of the LTER EML documents. It is noteworthy that we did not identify a single usage of the EML software module within the over 8000 EML documents provided by other ecological organisations such as the Partnership...
for Interdisciplinary Studies of Coastal Oceans (PISCO) and the Organisation of Biological Field Stations (OBFS). We surmise that software developers may prefer other established software project-documentation tools (like Sourceforge, The Sourceforge Software Repository, 1999) to describe their software. EML’s other two top modules, citation and protocol, are barely used at LTER (only three sites of the 26 made some use out of them). However, the EML community has proposed changes in EML’s citation module to improve fit with the bibliographical catalogues used at LTER. LTER believes that such changes will stimulate use of EML’s citation documentation feature.

Our analysis also revealed that EML database-related information placeholders (XML tags) are barely populated. LTER sites provide most of their data in tabular form, even in cases where the data are constructed via a query to a database. For example, the EML branch devoted to describe database’s stored procedures has never been used by any LTER sites. The EML branch storedProcedure shares the hierarchical level of data entity types like tabular data forms, spreadsheets, GIS data types (raster and vector types) and the EML database-related view element. This later EML element – view – which should be used to describe any custom queries (or views) of a database, is only used in one document.

Like the EML creators, the authors see some value for these particular database-oriented information elements, but the LTER experience indicates that these types are unused. For instance, the EML database-related element named constraint is barely used. The constraint is a complex element (a leaf or branch, an element that breaks down into other subordinate elements). The constraint set of elements describe integrity constraints within a relational database, including primary key constraints, foreign key constraints, and the like. Only eight LTER EML documents have used these constraint set.

Other widely unpopular elements in EML are the following: encodingMethod, authentication, series, compressionMethod, caseSensitive, connectionDefinition. Most of these EML elements belong in the leaf formed by the physical branch. This branch of EML covers detailed information about the physical characteristics of the container of the data. These EML elements, as well as others not cited here, were never used by any of the documents posted by LTER or OBFS. As in the case described in the last paragraph, all database-oriented descriptors have not been used by the LTER community. The EML series element seems to be a legacy element inherited from the Federal Geographic Data Committee (FGDC) metadata specification (FGDC, 2001). EML creators leveraged from other established specifications such as the FGDC, but unfortunately some of the elements inherited from other specifications serve little purpose for LTER. Encoding, authentication and compression methods are EML elements largely ignored by the LTER community. The same can be said about elements that describe concepts such as the maximum length for a record, the number of physical lines per record and the size of the data. Another element not used is offline, this is meant to describe data resources that are not available online.

Lastly, we are covering one additional element in EML that is seldom used by LTER sites. The element ‘describes’, under the additional metadata section, at the end of the schema, was only used by one LTER site. The specifications for EML suggest using this element to relate to other resource modules of EML that were not fully capable of accommodating content. In those cases the additional content could be placed under the EML additionalMetadata element and the original module would be referenced in describes. The example some EML developers had in hand was using describes and additionalMetadata to provide additional access rules for certain documents. However, the LTER community did not use describes at all, perhaps because we did not refer to other modules. The LTER community used additionalMetadata to provide metadata that did not fit well elsewhere in the EML schema instead. Another common use of the additionalMetadata placeholder was to provide detailed description of measurement units, following the Scientific, Technical and Medical Markup Language STMLML description (Murray-Rust and Rzepa, 2002) of the custom units used for the dataset. This practice and the previously described practices such as the measurement classification (Section 2), ignored are some of the common practices adopted de-facto within the LTER community.

3.2.4 Most commonly used EML elements in LTER

EML elements commonly used by the LTER community form a small subset of all the schema, yet, when coupled with wise Best Practices, formed a solid basis for data and metadata integration and homogeneity within the community. The following is a list of the common content one can expect when reading an LTER community, EML-compliant (controlled, structured metadata) dataset: A name, identifier with provenance information, version, title, abstract, personnel and institutional contact information, geo-temporal references, taxonomic information when applicable, a valid URL for accessing the data directly, notes on the methodology used, a thorough description of the data entity holder used (mostly, a table, or spreadsheet). The description of a physical datatable often includes the number of header lines, the labels of the columns along with descriptions and names, the units used, or range of values with pairs of codes used and corresponding definitions when applicable, missing data value codes used in each column or row, bounds for numerical values, date formats used, and field and line delimiters among other important descriptors to have the ability to understand the data completely, sometimes without human intervention (using intelligent systems such as those based on Provenance Aware Synthesis Tracking Architecture (PASTA) (Servilla et al., 2008)).
3.2.5 Additional features and information placeholders that add functionality to EML

Some information managers at LTER suggest that EML would benefit from the creation of additional information placeholders. Here are some examples:

Mathematical formulas: Currently, EML does not have an optimal mechanism to encode some mathematical expressions. Fractions, integrals, the infinite symbol and the summatory function are examples of important mathematical symbols that cannot be encoded in EML. The W3 consortium suggests using the MathML (Authors and Ausbrooks, 2003) specification for encoding mathematical functions in EML.

Scientific units: Another necessary EML area of improvement is the encoding of scientific units. MathML provides some standard formulation for scientific units. The LTER community formed a working group to improve the unit encoding disparity within LTER. The LTER Unit Task Force (LTER-UTF Units Task Force, 2008) is developing a vetting process for units within the LTER network as well as a custom unit dictionary.

Text markup: There has been some debate in the LTER community about the separation of content and form in EML. EML offers limited options for content presentation markup. The only content encoding options are new paragraph, text emphasis, ordered and enumerated lists, super and subscripts. LTER data managers reported critical format loss when encoding legacy metadata in EML. However, format-oriented markup tags must be prevented from interfering with the interpretation of the content itself.

Genomics-specific metadata: EML has been deemed inadequate to serve the metadata documentation needs of the genomics community (San Gil et al., 2008). The ecological and genomics communities have different approaches to data sharing, analysis and documentation. Ecologists have adopted general, sharable specifications while geneticists have developed specifications for each particular branch of genomics and each technology. Discipline-focused metadata specifications enable rapid data synthesis-oriented application development. The “Minimum information amount for a microarray experiment” (Brazma et al., 2001) metadata specification was among the first XML-based specifications designed to curtail the information entropy that menaced gene expression data integration. Many other genomics technology-tied metadata specifications followed the MIAME example, and guidelines to design these specifications have been created (Taylor, 2008). However, the practice of creating a standard for each group has been questioned (Ball, 2008). Recently, part of the EML community has been engaged with the Genomics Standards Consortium (GSC) in the process of designing a schema for genomics metadata (San Gil et al., 2008). The GSC has sponsored metadata language specifications (Field et al., 2008) focused on the documentation of genomic and metagenomic sequences. The LTER needs a genomics-community synergistic metadata specification for documenting LTER genomics studies. Extending EML using the additional metadata branch can effectively double the required work by presenting the metadata provider with two sets of forms. Merging the EML and GCDML specifications by adding a genomics report branch seems the path of less technological resistance. However, the differences in the specification focus addressed above present a problem that may impede the merging operation. As a solution for managing genomics data in their ecological context, the LTER is looking at extending the Sapelo Island Microbial Observatory environmental genomics implementation (Sheldon et al., 2002) to the LTER network level.

Quality control metadata enhancement: The LTER community has a working group on quality assurance and quality control processes (QA/QC). The working group concluded that EML does not provide sufficient structure for needed QA/QC descriptors. For example, some quality assurance observations are made at the record level (or individual observation level). The EML specification does not have descriptors at the individual record level.

4 Evaluating the content of EML metadata: measures of success for metadata providers

At LTER, we discovered that advanced functionality of metadata records, such as machine-enabled analysis, provides motivation to create EML documents beyond data preservation. We want to familiarise new EML users with LTER’s experience with tools to evaluate the functionality of filled EML documents. The many mechanisms to validate and review the efforts to create EML include the following:

A Check compliance with EML schema rules using a schema-validation tool. Many XML editors such as oXygen and XMLSpy provide this sort of functionality. In addition, the LTER data catalogue checks for schema-compliance at the time the EML document is stored. All EML document in the LTER data catalogue are schema compliant. Further integrity checks are made at the time an EML document is stored in the catalogue.

B Determine the potential functionality level of EML content per document. EML schema compliance does not indicate the potential functionality of the EML record. We have used simple Perl scripts to detect the presence of certain EML elements that indicate functionality. For example, the presence of a dataTable element indicates content-rich EML. In the absence of an entity-level element, we look for a data-resource
A few LTER sites have a thorough quality control. If we detect the presence of those elements, we catalogue EML functionality as discovery level.

C Assess quality control/quality assessment of the content. In step B, we show quick checks that assess metadata record functionality. However, the XML element detection algorithms do not evaluate the quality of the content. While there are several semi-automated tools that allow data consistencies to be checked based on the content, no tool can provide an assessment of the ultimate veracity of the associated data. A basic link-checker algorithm can attempt to retrieve online data using the content of the URL-distribution element. This simple check is sometimes prevented by authentication systems. When automatic data retrieval is possible, further automatic consistency checks can be done. The LTER Network Office co-developed a java-based data loader library that allows the automatic parsing of data based on the physical properties and attributes declared in EML. A successful consistency check occurs when the Data Loader creates a relational database table instance that mimics the original data structure. A similar consistency check for the associated data can be conducted using an online tool developed in Taiwan (Lin et al., 2007).

At LTER, we have used all the tools mentioned above. Based on these metadata evaluation checks, we can say the following about the LTER EML metadata (San Gil et al., 2009):

1. All LTER sites are contributing with discovery level EML to the LTER network data catalogue.
2. Most LTER sites (all but three sites) are contributing content-rich EML to the LTER network data catalogue.
3. A few LTER sites have a thorough quality control process for both data and metadata. Many LTER sites need to revise their EML for quality, as the template based EML implementation for legacy metadata called for a post-process manual revision. Also, some of the legacy metadata was not prepared for the advanced metadata functionality described here.

5 Conclusions
We re-examined scientific results previously published (Doran et al., 2002; Tilman and El Haddi, 1992) using a simple prototype driven by high quality, rich-content, structured metadata. We showed how the McMurdo Dry Valleys LTER area has experienced a cooling trend during the last two decades. We also corroborated Tilman’s conclusions on the negative effects of drought on plant diversity. We used a prototype of synthesis-application that leverages structured, high-quality, content-rich EML documents. The prototype is a simple combination of a Perl script data-digger and a custom Matlab visualisation script. Thus, we showed how metadata-driven data integration and analysis in ecology is possible when correctly using ad-hoc specifications such as the EML. We also provided a practical guide to use EML, as well as the LTER experience on using EML, including practical approaches to common EML usage challenges that were not entirely resolved by the EML guidelines or the EML Best Practices document. Finally, we proposed some functionality-oriented and user-experience enhancements for future revisions of the EML specification.

Acknowledgements
Inigo San Gil would like to acknowledge the support of the USGS National Biological Information Infrastructure through the NBII/LTER Cooperative Agreement.

References


Note

1Identifiers for EML documents that classify ‘Plot Number’ as: nominal (knb-lter-arc.1410.2), ordinal (knb-lter-fce.48.1), interval (knb-lter-ver.113.2) and even ratio (knb-lter-sev.88.2).