LTER EML Implementation Workshop

June 9 – 10, 2003
Sevilleta Research Station, Sevilleta National Wildlife Refuge, New Mexico

RESULTS/RECOMMENDATIONS OF LTER EML IMPLEMENTION WORKSHOP

July 25, 2003
NOTE: The various groups that contributed these recommendations have not had a chance to review them. The reports below are in raw form. (Sources documents: Best Practices, Software Strategies) Over the next 4 - 6 weeks, work will continue on refining this report.

- Agenda
- Tentative Working Groups
- EML Tiers Identified
- Best Practices: Beginning Steps Toward Developing a Set of Best Practices for Documenting Datasets Using EML.
- Strategies for Software Development: What practical steps can be taken in the short run to help LTER sites produce EML.
- Appendix 1: Technology Summary
- Appendix 2: XML Structure for EML documents for EML Tiers 1 - 5
- EML Handbook (Preliminary - not for public release)

Participants:
Peter McCartney (CAP*), Chris Jones (PISCO*), James Brunt (NET*), David Blankman (NET*), John Vande Castle (NET), Tim Bergsma (KBS*), Ken Ramsey (JRN*) Barbara Benson (NTL), Kristin Vanderbilt (SEV), Linda Powell (FCE)

*Member of the EML Development Group

Support:
NET Graduate Students: Jeanine McGann Co-author EML Handbook; Saurabh Sood and Gaurav Gupta, programmers of TextToEML Conversion Tools

---

Agenda

Monday June 9
7:00 – 8:00 Continental Breakfast on the Patio

Opening Group Session

8:00 – 8:15 **Define scope of workshop and expected products** (Brunt)

- Develop suggested metadata quality standards (best practices)
- Develop strategy for application development to handle mid-tier site EML implementation (relational, relational/XML combination, XML only).
- Strategy for bringing legacy metadata up to "best practices" standards.

8:15- 9:00 Report on Conversion Progress, tool development, and site plans for EML implementation (Blankman)

9:00 – 9:30 Report from the Network Information System Advisory Group (Henshaw) 30 minutes

- How EML fits in the NIS strategy?
- The concept of tiered information management capabilities.

9:30 – 9:45 BREAK

9:45 – 10:30 The possible tiers of EML Implementation (Blankman)

- Data Registry
- Data Catalog
- LTER Core
- The Whole Enchilada

10:30 – 11:00 Group discussion on how to proceed.

11:00 – 2:00 Working Group Session 1 (both groups same subject through lunch)
Process Managers: Group 1: Barbara Benson; Group 2: Kristin Vanderbilt
Purpose: Define a tiered EML implementation strategy

- Present "straw people"
- Each working group would discuss "straw people" example.

In this initial discussion we would be talking only about quantity not quality, that is, which modules/elements, should be considered in each tier – what functionality needs to be attained.

12:00 – 1:00 Lunch in the conference room

2:00 – 2:45 Groups present results of working group 1 discussions

2:45 – 3:00 BREAK

3:00 -> Working Group Session 2 (Monday Afternoon 3:00 PM ->)

- **Working Group 2A**. Process Manager: John Vande CastlePurpose: Developing qualitative "best practices" standards/recommendations for what good EML/metadata should look like – what quality metadata is needed to attain desired functionality.
- **Working Group 2B**.Process Manager: Linda PowellPurpose: Software development track. Developing strategies for application development for EML entry – how do we avoid developing 24 separate systems for entering and

editing basic metadata and making it available.

7:00 Van Leaves for Dinner at Socorro Springs

Tuesday June 10

7:00 – 8:00 Continental Breakfast on the Patio

8:00 – 9:00 Groups present results of working group 2 discussions

9:00 – 11:00 Work session 3 (2 groups)

- **Working Group 3A**. Process Manager: Ken Ramsey Purpose: Develop strategy for bringing existing legacy metadata up to "qualitative" and "quantitative" standards.
- **Working Group 3B**.Process Manager: Tim BergsmaPurpose: Software development track. Continuation of 2B with added focus of legacy conversion

11:00 – 12:00 Groups present results of working group 3 discussion.

12:00 – 1:00 LUNCH in the conference room

1:00 – 4:30 Writing assignments and small clean-up discussion groups

4:30 – 5:00 Final Wrap-up – What's left? – Who's Responsible?

Tentative Working Group Members

Group 1A	**Group 1B**
Process Manager	**Process Manager**
Barbara Benson	Kristin Vanderbilt
Members	**Members**
Chris Jones	David Blankman
James Brunt	Tim Bergsma
Ken Ramsey	Don Henshaw
Peter McCartney	Linda Powell
	John Vande Castle

Recorders/Support	**Recorders/Support**
Saurab Sood	Jeanine McGann

Best Practices	**Legacy Metadata Management Issues ***
Group 2A	**Group 3A**
Process Manager	**Process Manager**
John Vande Castle	Ken Ramsey
Members	**Members**
Barbara Benson	Kristin Vanderbilt
James Brunt	Don Henshaw
Kristin Vanderbilt	Barbara Benson
Don Henshaw	James Brunt
Recorders/Support	**Recorders/Support**
Jeanine McGann	Jeanine McGann
* The members of Group 2A continued their discussion of best practices. The software working group addressed some of the management issues related to converting legacy metadata.	

Software Devlopment	
Group 2B	**Group 3B**
Process Manager	**Process Manager**

Linda Powell	Tim Bergsma
Members	**Members**
Peter McCartney	Peter McCartney
Tim Bergsma	John Vande Castle
Chris Jones	Linda Powell
Ken Ramsey	Chris Jones
David Blankman	David Blankman
Recorders/Support	**Recorders/Support**
Saurab Sood	Gaurav Gupta
Gaurav Gupta	Saurab Sood

EML Tier Levels:

Following the concept of tiers as developed at the April 2003 Coordiating Committee the participants developed a a tier system for EML implementation by LTER sites.

Level	Description	Notes and Status
0:NO EML	Structured Legacy Metadata Text documentation to structured metadata but no use of EML	All LTER sites are, at minimum, at this stage
1:**Identfication**	Identification - Minimal Registry	All sites could be at this

	Information	level by coverting DTOC to EML
2: Discovery	Use Basic Resource Information	
2.5: Enhanced Discovery (Discovery+)	Level 2 plus coverage to enable spatial data, location discovery	
3: Evaluation	Use Core Candidate 2 excluding access but adding attribute, coverage, method and project/abstract	
4: Access	Use Core Candidate 3, adding physical attributes	
5: Integration (including QA/QC)	Use Core Candidate 3, adding constraint	All LTER sites should be developing the capacity to produce Level 5 EML.
6: Semantic Metadata	Currently under development by SEEK	Necessary for automated integration of datasets.

Best Practices

For the implementation the following was discussed as a best practices. Items that are included are suggested as most important for inclusion in the EML metadata, with other items left as optional.

For each module: each element

- Enumerate
- Describe content
- Workaround for existing EML structure (e.g. add methods, datum

Level 1 – Database identification

EML Resource group elements
< title> – Should be descriptive:
Needs to describe what and where
The dataset id for example AND0022, should be listed as the EML alternate identifier

(resource) and included in the minimal, level 1 requirement.

<creator> – full information for at least 1 creator – consistent format
< contact> – full contact should be kept current since this person is key in the long term. The contact must be kept current (NB: <contact> is not contained within the resource group, but it is located below the resource group module on the dataset level).

For <keywords>, core area and site name are needed as well as funding source (i.e. attribute to LTER, ILTER, co-funded with other sources, non-LTER funding etc.)

As an action item, an LTER Network-wide keyword thesaurus needs to be developed for use in the EML documentation.

The <abstract> will be useful for full-text search, and it should be rich with descriptive text. The measured parameters should be included.

<intellectualRights> should contain site data access policy, plus a description of any deviation from the general access policy specific for each particular datasets (ie restricted-access datasets). The timeframe for release should be included as well. For example, LTER Network-wide data should be released on-line within 2-3 years, and if not, the reason needs to be documented in the metadata.

Level 2 – Data discovery

For <distribution>, all options for eml-distribution can be included at the dataset level.
For <online> data, a <url> should be included which could point to the dataset, a script that generates a data stream, a data access system etc.
*automated access for data mechanisms to be developed
For <offline> data, the minimum that should be included is <mediumName>.

(NB: There should be an enhanced data discovery (a level 2.5) that includes coverage, so a search based on geographic area could be included).

Level 3 – Evaluation

<coverage>: In <geographicCoverage>, the method for determining <bounding Coordinates>, <boundingAltitudes>, coordinate datum etc. can be included under <geographicDescription> since it is a simple text field. The description should be a comprehensive description of the location including country, county or province, city, state, general topography, landmarks, rivers, etc.

Within <geographicCoverage>, <bounding Coordinates> should describe a rough bounding box (one point for each extension to the east, west, north, south) with the latitude and longitude to four decimal degrees in international convention (+_).
< boundingAltitudes> should be described in meters with a datum described in <altitudeUnits>. For bounding box location, polygon location information (<datasetGPolygon>), should be included when the bounding box does not adequately describe the study location. This needs further development.

For <temporalCoverage> representation, description of the date should be the date or <rangeOfDates> the database was collected (not the year the study was put together if it uses retrospective or historical data). Sometimes an <alternativeTimeScale> is more appropriate, such as the use of "years before present" for something like long-term tree ring chronology dating back hundreds of years. The date format should be listed as described in the EML documentation.

In general, it is recommended that <taxonomicCoverage> use the most specific scientific name wherever possible. The module also should include:
- <generalTaxonomicCoverage> – which should include a general textual description of all flora/fauna in the study, as well as how finely grained the taxonomy is broken down to – for example "family" or "genus and species."
- <taxonomicClassification> – scientific names are preferred to common names whenever possible.

The <methods> should be described at the dataset level and should contain the following:

- Under <methodStep> <description> <para> …. <literalLayout> should be selected to paste in text/html descriptions.
- Should describe exactly what the field crew did.
- Use <citation> for referral to other documents (use <distribution> <online> <url> within <citation> for referrals to online documents/manuals. (NB: the <title>, <creator>, and type of citation (<article>, <book>, etc. must also be included in this referral to produce a valid document.)
- If a widely used or published <protocol> is used, include that description
- For <instrumentation>, a full description of the instrument or instruments used should be included. Changes in instrumentation and dates of changes should be mentioned earlier in the methods description.

- Under <sampling>, use <studyExtent> to separate out more specific information about studies using simple text <description>.
- <samplingDescription> should be similar to a journal article sampling methods section. Use <literalLayout> to paste in text/html descriptions. Include sampling frequency.

The <project> information should include:
- <title> – Name of the LTER site
- <personnel> – Lead PI and information manager
- <abstract> - study area description and site DB or design description
- <citation> – Sites should create a citation for the home webpage of the site, using <distribution> <online> <url> (as described under <methodStep> best practices). For the resource group, should minimally link to the project or site URL home page, but could link to other publications describing the site.

Best practices for general <dataTable>, <entity>, and <attribute> descriptions include:

Entity group: <entityName> – title of the entity, such as the name of the table or file (needs clarification)

< entityDescription> – should be included if there are multiple entities and does not need to reinterate the dataset title

Each <attribute> should be the name of a field in a table
< attributeName> - if cryptic, consider using <attributeLabel> as well to provide a more intelligible name.
< attributeDefinition> – clear and complete; avoid cryptic definitions

<measurementScale>

For <nominal> or <ordinal> datasets:
Choose <enumeratedDomain> for use with coded variables, then provide <codeDefinition>

For <interval> or <ratio> datasets: <unit>, <precision> and <numericDomain> are all required.

If you have a <standardUnit>, use the name in the unit dictionary.
When constructing a <customUnit>, the units should be tracked, for instance in a common unit dictionary at the LTER Network Office.

Data Unit
Counts Number
Ratio
Percent PERCENT (custom use)
E.g. gmC/gm GM/GM

<precision>: use closest precision values. may need a work-around in some cases to describe data precision (when values in the same table have different precisions).

Within <numericDomain>:
< numberType> – should be defined as real, natural, whole or integer as explained in EML handbook
Bounds group – must understand that except for missing value codes, all data must be with in range, including outliers. May have just one of the <bounds>, either <minimum> or
< maximum>.

Other attribute-level modules:

<missingValueCode>: recommend against using missing value codes, but if used in data, it must be documented here.

There are no recommendations as yet for attribute-level <methods>.

Level 4 - Access

Includes <physical> module within the Entity group:

Required fields:
< objectName> - should represent the publicly available file with the specific file name (possibly exported as text from a database).
<dataFormat> -

Level 5 - Integration (including QA/QC)

Level 6 - Semantic Use

Software Strategies for EML Implementation

The choice of strategy for implementing EML in the LTER network is expected to be highly site-specific. Attempts to conceptualize and develop general solutions – suitable for seamless deployment at any LTER site – have not yet had the impact that was anticipated. In contrast, efforts to provide personnel and expertise to individual sites as needed have been highly successful (current status of available technical solutions are summarized in an Appendix below). These findings argue against investing resources exclusively in "one-size-fits-all" solutions to EML implementation. However, given the economies inherent in shared solutions, there is still justification for efforts to avoid inventing completely unique strategies for each LTER site. To reconcile the need for specificity with the economy of generality, we identify four site models, based on anecdotal knowledge of development constraints at known sites. For each of these models, we identify a generalized scenario for EML implementation, reviewing ramifications for metadata generation, formatting, editing, and sharing. Consideration is also given to model-specific strategies for conversion of legacy data and special maintenance concerns.

Model 1: Text-based system.

Some sites have an existing structured text-based metadata system. Generally this type of site has minimal information management staff. Metadata is provided primarily by researchers. Site IMs add physical file information, perhaps some other information. For this site, EML is viewed as a network participation product, and may or may not have a use for EML for other purposes.

For metadata generation, an entry template needs to be reviewed to make sure that EML-required metadata is provided. Any information such as attribute/variable tables needs to have consistent and unambiguous delimiters. Guidelines for need to be developed for know problem areas such as name and date entries. End product for EML conversion needs to be text (e.g.) site may use MS Word, but needs to have a version saved as text. For metadata formating and editing, NET will provide a site specific tool that work on single files or in batch mode to convert text-metadata from existing site metadata template to EML. Site IM will need to review eml and modify as needed.

For delivery and networking, this site type will most likely produce eml documents and publish them to a network-enabled archive such as a Metacat harvester system at NET.

For Legacy conversions, sites will need to provide units in canonical unit-dictionary form for valid eml. Most of the time legacy metadata will not coincide with unit-dictionary forms (e.g., grams/square meter not gramsPerMeterSquared). Quality issues will also need to be addressed. NET team will convert existing metadata. Site will be responsible for addressing quality and quantity issues.

Model 2: Creating a KNB/ecoGrid node using xml storage solutions.

For an LTER site wishing to publish its metadata to a shared search network standard such as KNB/ecoGrid, a strategy must address 1) solutions for creating and edting EML, 2) storing the EML data, and 3) articulating those data to some end user access application.

Metadata entry solutions would be those options designed to produce XML directly. These include text or available XML editors (Spy), a parsing strategy based on converting some alternate text format into EML, or using creation tools designed specifically for EML. As noted above no complete solutions exist within the later category, although morpho and xylographa offer some possibilities when combined with some post-process editing.

Storage options range from managing EML files in a file directory, or managing them in some native XML database solution. Avalable products range from Xindice (a public domain tool from apache.org) to Metacat (an xml storage system developed by NCEAS), to very pricey commercial solutions. Managing raw files of course requires merely knowledge of a file system. XML database products provide some interface for inserting and indexing documents either with a simple GUI, command line interface, or via an Application Programming Interface. A rich set of Metacat interfaces can be accessed via the Morpho application – currently this functionality is limited to beta versions of EML.

Providing Internet access to xml metadata can be accomplished by enabling linking your storage system to established search networks where client tools already exist or by developing a custom web interface. To accomplish the former: data stored in a metacat can be searched by a morpho instance that has been pointed to that archive. If your data are in a directory or in a Xindice system, you can install a Xanthoria target at your site which will expose your metadata to a xanthoria search client. To provide your own web interface, you could create a very simple static catalog by writing an html page listing each dataset, with a hyperlink to cause the chosen eml file to be displayed using an xsl stylsheet. Existing xsl stylesheets can be aquired from NCEAS or ASU as can some sample utilities for binding the xsl with your eml file. To provide search capabilities, the most logical approach is to write a custom search client that uses the client toolkit for Xanthoria or Metacat to provide search and display interface. Both ASU and NCEAS can provide sample applications, but either one is likely to require some customization as well as some configuration of required tools such as a java servlet engine or php processor.

Model 3: RDBMS systems.

Sites using (or planning to user) RDBMS for management of metadata will have somewhat of a challenge probably have workable mechanisms for capturing/uploading metadata. Although it is difficult to create relational designs that fully implement the elements of EML, useful subsets already exist at several sites. Usually these subsets are augmented with metadata categories that are site-specific, and can probably be augmented as resources allow with additional EML categories. Although requiring significant management expertise, RDBMS strategies enable a diversity of editing and output options. Various factors influence those choices made by the IM, including the current metadata content, technical expertise and resources, IM workload, and site culture (who generates metadata: PI, IM, techs).

Creation of metadata can follow an approach of (1) generating EML using one or more methods described above for editing or importing metadata and then loaded to the database or (2) using native entry editing tools. There are advantages to either approach but attempts to mix strategies are likely to become very compicated and lead to serious synchronization conflicts. Solutions for loading EML into an rdbms are limited to custom scripts, an experimental data loader for Xanthoria, Coccoon from Apache.org, and some vendor specific tools. All are very immature and require substantial customization or programming. A site with an existing database wanting to keep its existing metadata entry system will likely need to modify its schema and/or content to support the desired level of EML functionality.

Tools exist for harvesting RDBMS content for dynamic transformation to valid EML documents. Xanthoria, for instance, allows delivery of metadata content over a networked system with one-time configuration. . Proprietary solutions already deployed, such as ArcIMS Metadata Service, could potentially be wrapped with an EML- or ecoGrid-compliant search engine like Xanthoria, but this has not yet been done.Sites can also write or borrow/modify scripts and or style sheets for producing EML-formatted metadata locally. Most of the cost of EML implementation under model 3 arises from the long-term commitment to supporting RDBMS.

Each of the above scenarios would require a mechanism to harvest or ingest EML into either a local and/or network EML node. The node could be either Xanthoria which can connect the RDBMS directly to the search network or Metacat which would require a separate output solution to EML and then subsequent ingestion into Metacat..

Conclusion

While specific recommendations needed to be placed in the above contexts, the group did make some general suggestions. First, it was recommended that efforts by the Network office to visit sites and help develop in-situ solutions that help them arrive an stort term strategies for filling in some EML that leverage whatever local approaches they have should continue. Second, in order to shorten the time it takes to produce a "data registry" containing minimal EML2.0.0 for each site, the LTER network office should attempt to parse the metadata content generated out of the Data Table of Contents (DTOC), merge to the extent possible on the basis of site ID and creater last name with existing NIS dabases (SiteDB and PersDB) to generate which would at least provide a set of resource-level EML2.0.0 documents for all datasets currently in DTOC. These could be then distributed to the sites for either replacement with more complete documents or edited as "starter" documents using the above described methods.

Appendix 1: Technology summary

In order for site data managers to currently generate EML documents, a number of short-term strategies may be employed. These solutions range from very simple text editing, to community-built software tools designed to generate EML. These include:

A) Hand editing using a simple text editor, and validating the resultant EML file with an EML parser (http://knb.ecoinformatics.org/emlparser), or hand editing using a commercial XML editor such as XMLSpy (http://www.xmlspy.com).
B) Metadata conversion via a structured text parsing tool such as Data Junction or custom parsers developed by NET staff. Most of these solutions will require hand cleanup with some editor.
C) Metadata collection via Microsoft Excel-type templates (i.e http://gce-lter.marsci.uga.edu/lter/files/misc/GCE_data_subm_template.zip), with site specific java-based software tools that are developed at the LTER network office that translate the template into ASCII text, and then into EML encoded files. Validation as above.
D) Metadata entry via the Morpho data management tool, which currently produces the EML2.0.0beta6 version of EML, which may be 'exported' to EML2.0.0 via a menu command. These EML documents should then be reviewed with a text editor for conversion accuracy (i.e did the mesurement scales map correctly?) (http://knb.ecoinformatics.org/software/download.html#morpho)
E) Metadata creation via the reverse engineering of a relational database using tools such as Xylographa (http://es.asu.edu/bdi/Subjects/xylographa). This tool will create a partial EML document, that would then need to be finished in a text editor described above in order to create a valid EML2.0.0 file.
F) For spatially explicit data (i.e shapefiles), the ArcCatalog tool in ESRI's ArcGIS suite will create an FGDC profile document in XML, which may then be transformed to EML2.0.0 using an XSLT stylesheet (http://cvs.ecoinformatics.org/cvs/cvsweb.cgi/eml/lib/esri2eml/)

Appendix 2: LTER EML Levels

Links to sample xml files.

Note: Currently these only represent the structure of a document, i.e., no content.

Eventually these example files will be "best practices" candidates.

1:Identfication

2: Discovery

2.5: Enhanced Discovery (Discovery+)

3: Evaluation

4: Access

5: Integration (including QA/QC)

6: Semantic Metadata