LTER Information Management

Home

EML Best Practices Working Group

THIS WG's material has MOVED TO GIT.
See https://github.com/lter/IMC_WG_wired
contact Suzanne, Margaret or Marty

Membership:
Margaret O'Brien (chair), Corinna Gries, Emery Boose, Dan Bahauddin, Theresa Valentine, Ken Ramsey

The comment period is now closed (as of January 15, 2011)
The draft for review is available at: http://intranet.lternet.edu/im/files/im/emlbestpractices-2-20100901-DRAFT.pdf

Events in 2010:
2010 Sep 21-24: Focus topic of annual IMC meeting (KBS) MEETING PAGE | REPORT BELOW
2010 Jun 22-24: Production workshop at LNO to create a draft EVENT
Including Guest (day 1): Philip Tarrant

Title	Date Posted	Body
2010 May 27 VTC - BP Workshop Planning	05/26/2010 - 3:36pm	Participants: Margaret O'Brien (chair), Dan Bahauddin, Emery Boose, James Brunt, Corinna Gries, Don Henshaw, John Porter, Ken Ramsey, Inigo San Gil, Theresa Valentine See notes docs uploaded at the bottom of this page. **Workshop** Traveling to LNO: Corinna, Dan, Emery, Ken (driving), Theresa(?confirmed) and Margaret By VTC: Jonathan(?), Theresa(?) Interested lurkers: Wade, John Porter Agenda for today. What will we do, what will we cover, how will we publish results? Many changes are anticipated in the future. An active website could be used to publish results as a living document. But we also need specific versions to cite. Six years since last version. Scope of document. Write for EML 2.1. Reference old document for EML 2.0.1. Final document will be 20-25 pages total. Possible new data types include: GIS / spatial data, genomics, climDB and other specific projects. **GIS / Spatial Data** Search by point locations or bounding coordinates. This information appears in different places in EML (bounding box for entire site, points for plots, etc). Address both (1) how to use geographic coverage trees and (2) how to describe a spatial dataset. How to get spatial metadata into EML. A cross-walk tool is available for FGDC to EML. Handles attribute conversions, good editing tool in ArcGIS, but does not generate valid EML. Units need to be edited. Site-level polygons would be valuable for searches. Bounding box is not always appropriate. This may require a recommendation to eml-dev. Define levels of data compliance for spatial data? It may not be possible to address all issues at this time. **ClimDB** Recommendations for specific projects such as ClimDB? May require more than units, i.e. an attribute dictionary. ClimDB has a list of required attributes. But existing ClimDB metadata is quite specialized. Metadata in general are static and tend to become outdated. Find better ways to keep metadata up to date. Data provenance should ultimately be included in EML. E.g. data submitted to ClimDB should be accompanied by metadata that describe how the data were prepared. Common problem of what is data and what is metadata. Focus on helping sites create EML for data submissions to the network. **Data Models**

Title	Date Posted	Body
		Problem of different data models. Need a best practices recommendation. ODM data model. Long / skinny datasets (aka, "String of pearls"). E.g. record = site, date, parameter, value. This structure can be described in EML but the description is not useful. A processing step is required to create a matrix. Codes may appear in attribute description or in a separate table. How to describe a data table of this type? No place to describe units in EML. Create a new data type for EML? Or recommend against use of this data structure? Works for biodiversity data but not sensor data, etc. **Controlled Vocabulary** Duane is working with the HIVE project to include the current list and synonyms. HIVE has ability to enrich list. Tool will be available by end of June. Best practices. E.g. each EML document should have at least one keyword from the list, at least one core area, etc. The ability to identify a keyword thesaurus would be helpful. Establish a scope for each keyword (network, site, ad hoc). But not easy to encode at present. EML 2.1 supports an optional thesaurus tag (as a string value). How will keywords be used? What will a keyword search return? For now focus on getting lower-level terms (current list) into our EML documents. Higher-level terms (the tree) will develop over time. Users would pick keywords from a list. Only theme keywords? Other possibilities include: location, taxonomy, core area, decadal plan area, etc. Do core areas need to be updated? For search and browse, every dataset should link to at least one item, ideally far down the tree. But what about datasets that do not fit any of these terms? Synonym rings are critical to catch common variants, including variant spellings of the same word. Biomass or primary production, temperature or air temperature, etc. For now, use as many words from the list as possible. Unit dictionary can produce XML for inclusion in an EML file. A similar tool for controlled vocabulary would be useful. Grams carbon per meter squared or grams per meter squared (with carbon identified somewhere else). Try unit dictionary web services. **Wrap Up** IMC VTC June 7-8. Opportunity for anyone to comment on EML best practices. Margaret & Corinna will lead. Gather current practices and questions before the workshop. Search paths are indexed in MetaCat. Standardization of paths would help. How will conformance checker work with best practices? E.g. what constitutes an error, etc? How to manage the best practices document and updates? Maybe Google Docs. Goal is to produce a draft for IMC to consider in September. Publish in other locations as well? LTER website is good if focus is just LTER. Perhaps also ESA, Ecological Informatics, etc. Maybe write a paper. Workshop dates = June 22-24 morning. Travel on June 21 and June 24. Depart as late as possible on June 24. Margaret will send logistical details via email. Thinks about which sections you'd like to write.
Resources	05/27/2010 - 8:43am	2004 Best Practices document Link to PDF http://intranet.lternet.edu/im/files/im/emlbestpractices_oct2004_final.pdf http://intranet.lternet.edu/im/files/im/emlbestpractices_oct2004-3.doc (and attached below) IMC drupal page with general information about Best Practices and Guides. http://intranet.lternet.edu/im/im_practices/metadata/guides STORET codes http://www.dep.state.fl.us/labs/cgi-bin/wacs/codelist.asp http://www.epa.gov/storet/legacy/glossary.htm ONRL DAAC: Best Practices for Preparing Environmental Data Sets to Share and Archive http://daac.ornl.gov/cgi-bin/MDEDIT/bestprac.html SBC guides (audience: field and lab personnel creating datasets) http://intranet.lternet.edu/im/files/im/SBC_Attributes_Units_LTER_data_p...
2010 Jun 22-24 Workshop Logistics	05/27/2010 - 1:30pm	Logistics: See the logistics doc from George (attached below) there is a summary and map for Rapid Ride.
Representative EML docs from the LTER	06/10/2010 - 11:06am	These EML documents were submitted by LTER site Information Managers and are representative of the EML practices at their sites. Some may represent questionable practices. To view a document as XML, substitute the docid in this URL: http://metacat.lternet.edu/knb/metacat/knb-lter-and.4021.7 To view the document transformed to HTML: http://metacat.lternet.edu/knb/metacat/knb-lter-and.4021.7/default

Title	Date Posted	Body

AND
knb-lter-and.4021.7 (.8 contains html markup inside tags, but cdata tag was stripped by metacat. Here is correct version, with disable-output-escaping="yes" setting on "para":
http://sbc-dev.lternet.edu/cgi-bin/showDraftDataset.cgi?docid=CF002_exam...
knb-lter-and.4022.8
knb-lter-and.4341.16

ARC

BES
knb-lter-bes.240 - GIS data which could be expanded to a multi-geodatabase collection
knb-lter-bes.54 - ESRI GIS Geodatabase segment
knb-lter-bes.253 - ESRI GIS Geodatabase segment
knb-lter-bes.422.31 - Stream Chemistry Data.
knb-lter-bes.543.32 Bird survey file 1 of 4
knb-lter-bes.544.32 Bird survey file 2 of 4
knb-lter-bes.544.32 Bird survey file 3 of 4
knb-lter-bes.544.32 Bird survey file 4 of 4 These 4 are here as an example of eml currently at the discovery level which could be interesting to convert to level 5. The description within tells how to link the tables to reconstruct the relational database from the tab delimited text files. If that same information could be turned into attribute data in the proper nodes it should work.

BNZ
knb-lter-bnz.3.8
knb-lter-bnz.61.8
knb-lter-bnz.91.8
knb-lter-bnz.152.9
knb-lter-bnz.309.8
knb-lter-bnz.322.9
knb-lter-bnz.437.4

CCE
knb-lter-cce.15.2
knb-lter-cce.14.4
knb-lter-cce.153.2
knb-lter-cce.13.5
knb-lter-cce.17.2

CDR

CAP

CWT

FCE

GCE
knb-lter-gce.316.10 - nutrient data with lots of method sections
knb-lter-gce.320.9 - plant data set with simpler methods (data not yet public - data url function="information")
knb-lter-gce.284.11 - mooring time-series data with custom units
knb-lter-gce.215.12 - cruise ctd profile data with custom units
knb-lter-gce.290.11 - plant survey data with lots of taxonomy
knb-lter-gce.278.10 - soils data set with minimal methodology ("needs work" example)
knb-lter-gce.180.10 - fungi data set with questionable ('wacky') attributes/units ("needs work")
knb-lter-gce.1.14 - the data entity can be understood by John Porter's EML-R reader, but not by kepler. Could be the use of \r as line terminator.

HFR
knb-lter-hfr.1.12 - met data
knb-lter-hfr.2.11 - vegetation plot data
knb-lter-hfr.3.14 - phenology data
knb-lter-hfr.4.10 - eddy flux data
knb-lter-hfr.28.8 - paleo data
knb-lter-hfr.147.6 - ant species data

Title	Date Posted	Body

HBR

knb-lter-hbr.8.4 - Chemistry of Streamwater at HBEF WS-6
knb-lter-hbr.67.10 - Microbial biomass and activity
knb-lter-hbr.2.3 - Daily Streamflow by Watershed
knb-lter-hbr.53.2 - W5 Continuous Revegetation Survey Data

JRN

knb-lter-jrn.2002031.2 - Fixed width format defined for data table.

KBS

KNZ

LUQ

knb-lter-luq.21.2 - same methods and geographicCoverage repeated at entity level; temporalCoverage at entity level varies.

MCM

MCR

knb-lter-mcr.5.15 - redundant geog and time coverage (dataset and entity levels). methods at dataset level but should be (also?) at entity level.
knb-lter-mcr.6.36 - taxonomicCoverage only to family because 400 species not practical for this format. Complete taxonomy listed in separate entity. Example of primaryKey constraintType.
knb-lter-mcr.5001.4 - inline data, perhaps better as enumeration? (mesh table)
knb-lter-mcr.5003.3 - example of otherEntity
knb-lter-mcr.4001.5 - example of primaryKey, foreignKey and joinCondition. Header and footer line. Enumerated domains. Inline data.
All mcr datasets require a human to fill out a form prior to data delivery. Yet PASTA or the ECC can still access the data directly because those IP addresses are "white listed".

NTL

no machine access to the actual data files yet.

NWT

knb-lter-nwt.414.6 - daily precipitation, C1 climate station
knb-lter-nwt.411.5 - daily temperature, C1 climate station
knb-lter-nwt.105.6 - Green Lake 4 streamflow
knb-lter-nwt.108.6 - Green Lake 4 stream water quality
knb-lter-nwt.31.9 - snow survey data
knb-lter-nwt.96.9 - snow water equivalent data
knb-lter-nwt.419.3 - aboveground plant biomass data
knb-lter-nwt.407.2 - plant species composition data

PAL

knb-lter-pal.33.3
knb-lter-pal.28.3

PIE

knb-lter-pie.1.8 - Rates of benthic metabolism and nutrient cycling
knb-lter-pie.139.2 - Phytoplankton identification using HPLC and Chem Taxonomy along transects
knb-lter-pie.3.6 - Water-column nutrient and particulate transects
knb-lter-pie.4.5 - Dawn/dusk water column data along transects
knb-lter-pie.8.3 - Monthly time series of macrofauna sampling
knb-lter-pie.28.4 - Marsh plant biomass data
knb-lter-pie.13.4 - Carbon and nitrogen isotope data by year by functional group by site
knb-lter-pie.67.4- Year 2007 weather station data, 15 minute intervals
knb-lter-pie.136.2 - Annual Nutrient Loading and Yield
knb-lter-pie.106.4 - Nutrient data between 2001 and 2009 in three headwater sites

SBC

knb-lter-sbc.10.15 - two tables: CTD + bottle profiles; many protocols as pdfs w/urls all at the dataset level, <creator> has a <userId> child. Uses <references> in place of multiple trees.
knb-lter-sbc.17.17 - time series of fish counts, ambient abundances, compare to dataset #30. Attribute site uses enumerated list, redundant geoCov (ref not available)
knb-lter-sbc.30.4 - time series of fish counts from a kelp-removal experiment. nearly identical table to #17. What content helps to distinguish these?
SBCLTER_streamDischarge_TEMPLATE.1 - example of template used for many identically formatted tables. not public. non-standard packageId format, and no data table.

SEV

knb-lter-sev.21407.1 - Ground-truthing Satellite Imagery with Phenological Observations at the Sevilleta NWR -- Visual Observations (reuse of geographic coverage; it's just a duplicate entered by software that generates EML)

Title	Date Posted	Body
		knb-lter-sev.065.1 -- This is not a great example, but what I would appreciate is a better understanding of how to use the and fields. How does studyExtent differ from GeographicCoverage and how does differ from a containing a description of how the data were collected? **SGS** knb-lter-sgs.1.1 - SGS-LTER Long Term Monitoring Program: Spotlight Rabbit Count (http://sgslter.colostate.edu/Data/EML/LTMntrRabbitCount.xml) knb-lter-sgs.56.1 - CO2 Elevation Study: Biomass by Species, from ambient and elevated CO2 OTCs and unchambered controls (http://sgslter.colostate.edu/Data/EML/otc_sppharvest.xml) - this metadata refers to a plant list, which would need to be obtained from the researcher. knb-lter-sgs.18.1 - Ecosystem Stress Area: Long-term density data following nutrient enrichment stress (http://sgslter.colostate.edu/Data/EML/esa_den.xml) - this metadata defines the species codes, but they are not standardized or modernized according to any authority, but rather keep the same historical site codes throughout the long-term range of the data. **VCR** knb-lter-vcr.49.8 - column-formatted data with lots of codes knb-lter-vcr.167.1 - comma-delimited data from lots of locations knb-lter-vcr.54.4 - a spreadsheet-based model where the model formulae are in the spreadsheet knb-lter-vcr.148.5 - dataset with minimal internal documentation, but lots of links to external documents including videos knb-lter-vcr.89.5 - metadata includes statistical program used to summarize the data knb-lter-vcr.145.2 - GIS vector data with FGDC metadata encapsulated in the Abstract/general methods ??? knb-lter-vcr.172.1 - Bathymetry data in multiple raster GIS formats but not using the spatial module of EML ???
2010 Jun 7/8: VTC Watercooler Notes - IMC input on current Practices	06/11/2010 - 9:48am	MONDAY Participants: Margaret (lead), Dan, Don, Emery, Gastil, James M., Jason, Kristin, Mark, Mason, Suzanne The topics we touched on: 1 there should be a general resource for anyone using EML, i,e, beyond the normative docs, but not specific to LTER. examples of common needs: a) where are all the places for coverage trees. how are these best used? b) what are all the elements that have the attribute group id, system and scope? (LTER recommendation might be how to phrase content for the system attribute) c) which fields are intended to be machine readable, vs human? example, methods is not likely to be machine parsed. 2. What are the current modes for constructing EML in the LTER? [I recall Inigo doing something like this during EML2.1 development for LTER docs. ie, he examined a lot of XPATHs and reported on which were not used] before the workshop ask each site to give us the IDs of one or two "representative" docs showing where they put certain items. coverage trees have lots of locations - what are the most common locations, why? what kind of bounding boxes are included (whole site? edges of sampling area?) where do they put methods do you try to package up zip files? (eg, batches of GIS data) 3. Metadata Levels Current BP shows 5 levels (1:5). these could be reduced to 1:3. But then use levels 4 and 5 for datasets which have data attached e.g., L4=human readable, L5=machine readable (congruent-passed, or intended to be passed) 4. Congruency checker: a) create a separate section in BP-doc listing fields which congruency checker will examine, and what is expectation b) congruency checker should report on content the BP-doc recommends for levels 1:3 also. 5. EML trees that have inadequate recommendations include: a) methods/sampling, samplingExtent, spatialSamplingUnits b) coverage/taxonomy [keep in mind that there are no species in the controlled vocab. so metacat will need to search taxa] c) dataTable/*/attribute [mob added, not discussed] 6. A theme that cropped up several times was to identify use cases: a). where do scientists look to find the info they need in a dataset? e.g., do they need taxonomic family? just species binomial? b). geographicCoverage: where does LTERMaps want to look? LTER could suggest which datum to use for fields that do not have a datum specified. this would be for the general mapping fields - geographicCOverage c). [from gastil and mob, not discussed: search paths used by metacat are indexed. These need to be published and ref'd in BP] 7. Possible use of the "EML breakout" at the annual meeting. [We had originally put EML on the agenda because we thought we might have some output from a beta-congruency checker to report. But that will not be the case.] These ideas cropped up, though, and could be made product-oriented, and very informative. a). Break into groups and discuss/compare your site's EML with other IMs. Bring particularly difficult-to-describe or unusual data (although this might not be as useful as the norm). Get feedback from others, compare structures and possible uses. This was suggested by 2 people who have found this experience useful at their local sites. Dubbed "The doctor is IN" by Mark Servilla.

Title	Date Posted	Body

b). Breakouts could be centered around science themes. Examine current EML contributed to the NIS for those themes, how these are constructed and comment on which structures worked best. This format could help identify the "use cases" listed above (item 6a). Compare the docs to the best practices draft.

MORE FROM MONDAY

EML best practices group has been reassembled with a workshop later this month. Plan to address EML 2.1, spatial data, unit dictionary, controlled vocabulary. Each dictionary may have its own best practices.

EML sampling. How should this element be used?

EML taxonomy. What is best approach? Look at different sites to see range of usage.

EML keywords / thesauri. Lots of variability among sites.

Which are the key fields for the data congruency checker?

EML five levels of compliance. Are these still useful? Reduce to three? Or modify (e.g. does metadata support automatic data loading)? Possible levels include: Identification (1), evaluation (3), integration / validated as pasta-ready (4-5).

EML best practices should focus on what goes into the EML document. Congruency checking is a different (but related) issue.

How to check for metadata completeness? Completeness may be sufficient for levels 1-3. Congruency checker will focus on agreement between data and metadata. Pasta can check that required fields are filled out (but not check their semantics).

EML spatial bounding coordinates. ClimDB is missing detailed spatial coordinates for individual stations. Need recommendation on which coordinates to store and where to store them. EML provides many options. How to use EML to populate ClimDB, etc?

Workshop could look at different use cases. Some comments to date from the LTER Maps group.

LTER taxonomic coverage. Do users need only genus & species or entire taxonomic tree? How will users do taxonomic searches? LTER Metacat advanced option provides some search capability.

ITIS web service. No link from current EML taxonomic coverage. More generally: which elements have an external service (e.g. ID)? This will become more important as the number of online databases increases and may provide additional checking in the future. Develop use of ID in future versions of EML. This would be a parser (not schema) change. At present EML system attribute is largely ignored.

EML coverage elements are general (for Google maps, etc). Spatial datum is found only at lower level. Add recommendation for preferred network datum (e.g. WS1984)?

Some items (e.g. system identifiers) may be appropriate for the ecoinformatics community. Still room for LTER Network to identify preferred external systems.

Create a mechanism for human review of sample EML files (esp. for beginners)? Start a tradition of pairing up or assembling small teams at IMC meeting, etc. Just reading the normative documents has its limitations.

Eml-dev is also a resource. No problem with asking questions about EML implementation. Often not many responders but a good place to get expert feedback.

Current EML best practices document was effective at anticipating many problems. Examination of EML documents across the network would highlight current problems and concerns.

MORE FROM TUESDAY
John Porter, linda powell:, hap, sven, mark servilla/yang, jonathan, wade, theresa, nicole, corinna, Duane

examples are good. keep them or expand.
links to svn in existing doc are broken.
Should vs must. e.g., what are recommendations vs what is required by checker.

methods: should these be parseable? should there be a contact? come up with a recommendation for what is expectation for each use of methods. eg, at the attribute level, could help define the attribute, where as at dataset level could describe sampling.
also get info here on how data are intended to be aggregated. assist in decisions to reuse.
recommend including instrumentation
can we devise a way to refer to a standard method? (does methods tag have id/system/scope?)
we need more tools. for instance, an attribute dictionary could also include (or export) a methods tree as part of the description.
what is possible based on realistic practice? examine current EML files in this area. what do we do if we get vastly varied approaches? should they be consolidated into one?
machine vs human readable: some sections (like methods) are probably not going to be machine readable.

views vs DB tables:
dataTables that have been denormalized, but we would like to put back into a normalized form. If you produce both, then put both tables into one dataset. Good place for a common tool -- produce a view and a normalized table.

Notes from Tuesday, June 8, 2010

Title	Date Posted	Body

IM_VTC_EML_Best_Prac_20100608

in attendance: duane, Nicole, Margaret, hap, jonathan, mark, yang, linda, john, sven, wade, thersa, corinna

• what needs to be changed?

o authoring section - what if someone leaves

ƒ one of reasons we suggest listing IM office as contact

ƒ keeping contact info current is hard! on levels - need to be more specific and focus on quality

ƒ previous document had tiered trajectory as heavy influence

ƒ some aspects are arbitrary - little middle ground

ƒ we should back off on the levels some

ƒ focus on important drivers for machine usability

ƒ now tends to be binary - level 5 or not

ƒ like to see more on quality

ƒ keywords, units, attributes

ƒ did not do much on content standardizatoin in previous document

ƒ think more about content, less about structure content standardization should this group take on?

ƒ yes - if not here, where?

ƒ can use "must" "should" to delineate degree of concensus

ƒ issue - how methods are entered

ƒ very few give specific method steps

ƒ we may decide its OK to aggregate

ƒ 3 different places for methods in EML

ƒ human readable

ƒ but can be displayed selectively with tools if marked up appropriately

ƒ guidelines for content are a good way to go.... process look to see what sites do...

ƒ but what should we do where there is wide variance

ƒ can recognize variance and discuss to come up with recommendations

ƒ locations of research sites as metadata or as data

ƒ EML coords as bounding box or down to sample level?

ƒ currently recommend bounding box for areas with samples

ƒ probably "should" not "must" if disagreement

GIS data - issues of FGDC vs EML metadata

ƒ many sites not using EML for GIS data

ƒ locations as data vs metadata

ƒ can see how original recommendations map to future uses

ƒ ESRI metadata and EML

ƒ describing tables in EML

ƒ is level 5 EML for GIS data really useful beyond discovery-level

ƒ could use in PASTA

ƒ can use in KEPLER for GIS data

ƒ attributes in DBF file

ƒ metadata is part of shapefile - so you get nested metadata - FGDC inside EML

ƒ want to get FGDC translator working better - then can pull out FGDC data into EML easily

ƒ can include attributes, ranges, domains, codes etc.

ƒ FGDC lacks units etc. that are needed for EML data integration

ƒ methods are important

ƒ experimental design

ƒ sampling unit or unit of analysis

ƒ want to include this as specifically as possible

ƒ different locations for methods

ƒ different scopes

ƒ ties into derived products -may only want methods for some attributes

ƒ getting methods in any form is difficult!

ƒ aren't there standard methods? could those be incorporated? machine readable needed for PASTA

ƒ Data Manager Library tests for "pasta ready"

ƒ does the metadata match the data?

ƒ if it doesn't, can't load data

ƒ in reviewing best practices, at level 3 there is a lot to make a doc PASTA-ready

Title	Date Posted	Body
		ƒ could also include other quality checks ƒ but won't be restricting harvest - just reporting would be nice to get normalized forms of tables - not just denormalized forms for input back into a database ƒ do I want to offer normalized data ƒ or share denormalized ƒ I'd suggest both ƒ try to develop tools for doing that we can share examples were really helpful ƒ had fun doing those last time! ƒ all of old links are now broken when SVN was refactored attributes and chemical units - units of WHAT ƒ right now attributes are text where can we get periodic table data in here attribute standardization? for right with EML structure now need to have element name in attribute description could use kludge of setting up standard units that include objects of the units - with basic units as link to them in unit registry would be needed for unit conversions in PASTA ƒ but that is not planned to be automated ƒ robust descriptions are enough EPA Storet codes might be a good way to go ƒ links methods, units, objects together ƒ lots of opinions on STORET • next steps 1. o want you to list several representative EML IDs from your site 2. o look to see what metadata standards other EONs are using? ƒ good idea, but sometimes hard to do ƒ NEON may be using EML ƒ may be broader issue

Title	Date Posted	Body
2010 Sep 23 IMC meeting (KBS) EML BP Breakout	10/03/2010 - 8:19am	Notes assembled from IMC on EML BP doc content and organization notes from breakout group 2 (Corinna) Note takers: Mason Cortz and John Porter EML Best Practices Feedback 1) 'as granular as possible' should be 'as granular as necessary' EML Congruency Checker General Comments • Checker should report errors but continue when possible (no fail-on-first-error) • Checker should include a data file row number for each error • Errors should include a brief description of the tag/entity that is in error (user friendliness for users who are less familiar with the EML schema) • EML levels should not be used by the checker but importance for o needed for discovery o needed for use o needed to assess suitability for use • Start with CSV files Checker Features 1) Link to data is present and live a. Option to provide data file/URL and continue (congruency checker) b. Location should follow EML best practices 2) File format must be as described in data table and parsable a. Priority is for text formats b. Secondary concerns are externally defined formats 3) Number of columns defined in EML match the number of variables in file a. Defines which rows have the wrong column count (if not all rows) 4) Match EML columns to file columns by order and/or header names a. Check ordering first, then match to headers or fall back to ordering if no header is defined (whether ordering or name-matching takes precedence should be addressed in EML best practices) b. Have to define numHeaderLines, recordDelimter, and physicalLineDelimeter for this feature 5) Check that storage type in EML matches assumed type from data file a. Defined missing values that are not in the specified storage type are still okay (e.g. NA is okay in a numeric column if it has been defined in the EML as a missing value) 6) Check that the ranges in the EML document conform to the storage type 7) Check that the values in the file fall within the ranges a. Exceptions made for defined missing values 8) Check the date/time in file matches format string defined in EML attribute 9) Check that values for code columns are defined in codeDefinition a. Future feature: check external code sets if given in parsable format 10) Check for existence of fields required by the EML best practices but not by schema parser a. Is a field 'important'? Does it help with 1) discovery 2) ingestion/use or 3) explaining suitability for particular use e.g. check for geographic, taxonomic coverage, pud date, methods b. Error if fields are filled out in one entity but not another (e.g. geographic coverage present for organization but not for project) c. Return a report on optional fields that are not present in the EML document d. Future feature: check against NetworkDB for units, controlled vocabularies, personnel

Title	Date Posted	Body
2010 Sep 21-24: EML Best Practices WG Report from 2010 IMC meeting (KBS)	10/26/2010 - 10:00pm	The following is primarily a summary of discussions in the WG at the IMC annual meeting at KBS (September 2010). Links to other pertinent notes and reports are in Related Materials section. The following table shows a grid created during 2010 IMC Meeting (KBS), which was suggested to replace the "EML Metadata Levels" in previous versions of the EML Best Practice Recommendation. Whereas the earlier levels described only metadata, this view includes stages of data/metadata usability. It was intended to convey that although any schema-valid data package can be contributed, only packages that can be ingested into PASTA can be used for synthesis. Table 1 reflects data as it advances from Level 0 (at sites) to Level 1 (in the NIS data cache), and not the usability of data for synthesis. Possibly this additional state could be included by expanding the table to the right and/or down. In this context, "data usability" means ingestion into the PASTA "data cache", which is assumed to be a relational DB table. The current "catalog" is Metacat. Boxes in the left column refer to metadata, and on the right to data, and usability increases to the right, and down. | | Data Usability > | | Metadata Usability \V | | Schema Valid Examples: a. no dataTable node, minimal EML content (old "Identification" and "Discovery", levels) b. Data cannot be read, EML content varies | Undocumented, clean data Examples: a. dataTable can be ingested to data cache, but has minimal EML content b. measurementScale set to "nominal" with apparently numeric data (TDB – this is actually 'poorly documented' data) | | | Catalog Ready Examples: a. certain pre-determined elements are present and filled in (list TBD) b dataTable can be read, but not ingested to cache. EML quality undetermined. c. URL present and readable, but otherEntity with no attributeList | PASTA Ready Examples: a. complete EML, data can be ingested to data cache b. uses LTER terms and attribute features (e.g., units) | The EML Best Practices Group has determined list of action items and schedule: 1. Jan 31 2011: Create ToR (pending completion of the template by the Governance WG) 2. 15 Jan 2011: Comments due from IMC 3. 15 Feb 2011: Document available as PDF on IMC website 4.15 Apr 2011: Final Review due from IMC 5. 1 June 2011: Publish HTML version on IMC website
Representing EML datasets in the ORNL DAAC	12/10/2010 - 12:42pm	LTER datasets are contributed to the ORNL DAAC. CONTRIBUTION PROCESS: Process and frequency are unknown. TO DO: Find a link to documentation. RESULT: The representation of the dataset originally contributed to LTER is represented a little differently in the DAAC, and sometimes not as intended. Example: this is knb-lter-sbc.6.2 http://mercury.ornl.gov/ornldaac/send/xsltText2?fileURL=d:\mercury_instances\ornldaac\lter\harvested\metacat.lternet.edu_knb_metacat_action_read_qformat_xml_docid_knb-lter-sbc.6.2_1.xml&full_datasource=LTER%20Data&full_queryString=%20text%20:%20melack%20AND%20%28%20datasource%20:%28%20daac%20landval%20rgd%20lpcol%20lter%20obfs%20%20%29%20%29%20&ds_id= Notes: 1. This link was visited in Dec 2010, but this is not the most current version of this dataset (current revision is knb-lter-sbc.6.6). So update frequency is unknown 2. there were 2 data entities in knb-lter-sbc.6.2, but these have been decoupled in the DAAC view. One data entity was metadata (site locations), not "chemistry data" at all. 3. the URLs in the DAAC view were interpreted in the opposite fashion from what was intended. - The "Data Center" label is attached to the entity download URL, which in SBC's datasets is at the entity/physical level - The "Download Data Sets" label is with the SBC url, which SBC puts at the resource level with the attribute function="information". (Possibly, in rev 2, these may have been function="download" because early datasets were built with morpho, which used the default attribute value.) QUESTIONS (EML-related): 1. Are these dataset displays transformed from EML? or from EML that has been transformed first to FGDC? If FGDC, does LTER control the transformation? 2. should we recommend that certain elements be used to standardize their view in the DAAC? 3. How should transforms for (or by) the DAAC handle the occurrence of multiple entities/dataset?

Title	Date Posted	Body
Comments on EML Best Practices Draft, January 2011	01/04/2011 - 10:45am	This page is to be used for gathering comments on the EML Best Practices Draft for the Best Practices group to use in their final version. You may either add your comments to this page (as a comment) or add an entire document, which might be the draft with your comments included using "track changes". How to Add a comment: Click "Add New Comment" at the bottom of the page (you will be redirected to a login page) How to attach a document: log in click "Edit" (top of page) click "File Attachments" (scroll down to near bottom) browse to a local document with your comments, Click "Attach". Please include your name in the file label. click "Submit"

Title	Date Posted	Body
Metadata Content Guide for Scientists	09/10/2012 - 5:52am	This is a preliminary page where we will collect either links or attachments to existing guides from sites who have been using these with success to guide their scientists and graduate students. The guides specify what metadata content to include and may include examples. **AND** Overall: http://andrewsforest.oregonstate.edu/data/metadata.cfm?topnav=115#desc Dataset-level Metadata (includes entity-level definitions): http://and.lternet.edu/lter/data/metadata/metadata_desc.doc Entity-level Metadata: http://andrewsforest.oregonstate.edu/data/metadata/attribute_table.xls **BES** The BES template is an Access database form. It will include attribute information soon. See the attachment titled "bes-metadata-example.jpg" for an image. **CAP** See Attachment CAP_metadata_form_V2.doc **CWT** Coweeta's metadata guide: http://coweetagis.anthro.uga.edu/cwt_kb/index.php?CategoryID=41 is a set of web pages, including a zip of Data Submissions Forms (see link on Intro page). **FCE** See attachment below of Excel spreadsheet template xls_eml_01_FCE.xls. The guide is embedded within the cell heading. Note the eml generated from this template is still the EML 2.0. (For later version see ARC's) **GCE** GCE Data Submission training presentation and example files for investigators, technicians and graduate students: http://gce.lternet.edu/public/app/resource_details.asp?id=101 See attachment GCE_Data_Submission_Template_Processing.pdf (below) for GCE IM Office protocols for reviewing and post-processing submissions. A generalized version of the Excel template included with the training materials and post-processing instructions will be included in the next release of the GCE Data Toolbox software. **KBS** KBS uses Submitting_Data_to_the_KBS_LTER_Data_Catalog.doc to help KBS LTER folks know how to prepare their files and metadata. (Attached) **LUQ** Rules, protocols related to data files and templates to provide metadata to be posted at the LUQ IMS-website: http://luq.lternet.edu/IM/rulesProtocolsandTemplatesforDataFiling **MCR** See attachment named MCR_RAPID_template.xlsx. This is not a sophisticated Excel to EML thing; it is just an uncontrolled spreadsheet. One of our investigators made it for himself. **NWT** See attached Excel spreadsheet NWT_metadata_and_data_template.xlsx, which was put together by NWT's previous IM. It does not produce EML, but we can generate the ascii-format files we use for our posted metadata and data by cutting and pasting. **PIE** See attached Excel metadata template developed by ARC and modified for PIE. File is also available at http://ecosystems.mbl.edu/pie/data/protocol/PIEMetadataBlank.xls. Cell comments provide a guide for cell entry. A separate macro also developed by ARC is used to generate EML 2.1.0. **SBC** 1. See attachment named SBC_metadata_submission_template.txt. This is a plain text file, based on the GCE template. 2. SBC also uses the template developed by MCR 3. SBC's process for adding data packages to the catalog is in attachment named SBC_data_publication_checklist_redacted.txt. This is a plain text file. 4. SBC has a guide to attributes and units for data contributors (e.g., scientists or techs). It summarizes info from many other sources including the UnitsDB best practices. See: http://sbc.lternet.edu/external/InformationManagement/documents/SBC/SBC_Attributes_Units_LTER_data_packages.pdf 5. We also have an old Morpho-guide for scientists which could be generalized. **VCR** The VCR/LTER uses a web-based form system for input of metadata. A (rather ugly) document showing some sample forms and the underlying schema for the database is located HERE. Screencast videos of using the system are available HERE. However, be warned these are in .avi format and quite large. A powerpoint (with sound) on preparing data for submission is available HERE. Additional help (linked to from the form-based system) is at: http://www.vcrlter.virginia.edu/data/submission.html.

This material is based upon work supported by the National Science Foundation under Cooperative Agreement #DEB-0236154. Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s)