Jornada Metadata Standards for EML Creation

The Information Management Team

2019-10-08

Table of Contents

[Introduction 2](#_Toc21431320)

[Using this document 2](#_Toc21431321)

[Metadata standards for EML <dataset> elements 2](#_Toc21431322)

[<title>\* 2](#_Toc21431323)

[Personnel and organization elements (<creator>, <project>, etc.) 3](#_Toc21431324)

[<contact>\* 3](#_Toc21431325)

[<creator>\* 3](#_Toc21431326)

[<project> 4](#_Toc21431327)

[Other personnel and organization elements 4](#_Toc21431328)

[<pubDate> 4](#_Toc21431329)

[<abstract> 4](#_Toc21431330)

[<keywordSet> 5](#_Toc21431331)

[<intellectualRights> 5](#_Toc21431332)

[<coverage> 6](#_Toc21431333)

[<geographicCoverage> 6](#_Toc21431334)

[<taxonomicCoverage> 6](#_Toc21431335)

[<temporalCoverage> 7](#_Toc21431336)

[<maintenance> 7](#_Toc21431337)

[<methods> 7](#_Toc21431338)

[Data entities 8](#_Toc21431339)

[Information management, metadata, and EML resources 8](#_Toc21431340)

**If you edit this document please track your changes and send to Greg Maurer (gmaurer.jrn.lter@gmail.com).**

**Notes about things that need to be expanded or verified are marked with “???”**

# Introduction

This document defines metadata standards to use when creating or revising EML files for Jornada Basin LTER research data packages. These standards are under active development as we revamp our data catalog, but they are essentially derived from this reference document:

<https://environmentaldatainitiative.files.wordpress.com/2017/11/emlbestpractices-v3.pdf>

with modifications to suit the Jornada Basin LTER and its researchers.

When revising earlier EML documents, do your best not to lose the metadata they contain. Make sure a copy is archived somewhere (like EDI, if possible) so that metadata are not lost.

## Using this document

The metadata standards in this document are structured and formatted similarly to an EML document. Named EML elements are surrounded by angle brackets, and there are headings for most important elements. At the Jornada Basin LTER, a variety of tools are being used to create EML documents during the data packaging process (EMLassemblyline, GCE Toolbox +, etc.). Notes about using these tools to create each EML element are present in the relevant section of this document (???**Add these in as you edit**).

Note that EML documents are hierarchical. The root element is <eml:eml> (denoted by the path /eml:eml). Below the EML root element are 3 elements, <access>, <dataset>, and <additional\_metadata> (paths /eml:eml/access, /eml:eml/dataset, and /eml:eml/additional\_metadata). This document contains metadata standards for elements under the <dataset> element of Jornada EML files only. We may add standards for EML <access> and <additional\_metadata> elements at a later time.

# Metadata standards for EML <dataset> elements

The sections below correspond to elements under the <dataset> element of a Jornada Basin LTER EML document. They appear in roughly the order they should in a real EML document.

## <title>\*

**\* This is a required element.**

The <title> element should contain a descriptive title that includes type of data collected, geographic indicator, and temporal range of the data (what, where, when). **Keep in mind that package titles are searchable and are a user’s first point of contact when searching for data packages.**

Best practices:

* For the title’s geographic indicator all Jornada Basin LTER data packages should include “Jornada Basin LTER”.
* If the data come from an established study with a name or acronym, include the name or acronym in the title (NPP study, SMES, NEAT, etc.)
* For the temporal range of completed data packages give the starting year of the data to the ending year; “2009-2015”, for example.
* For the temporal range of ongoing data packages there are 2 possibilities:
	+ Write the end of the temporal range as a year if data collection occurs less frequently than yearly (“2010-2018” for a package collected every 2 years).
	+ Write the end of the temporal range as “ongoing” if data collection occurs yearly or more frequently (“2010-ongoing” if data were last updated a year ago or less).

## Personnel and organization elements (<creator>, <project>, etc.)

There are a number of elements in this category grouped together here. They normally appear within the <dataset> element (/eml:eml/dataset/creator, /eml:eml/dataset/contact, etc)

At least one <contact> element must be supplied in every EML document. All EML documents should have at least one person in a <creator> element, but exceptions can be made if this information has been lost in very old Jornada data packages.

### <contact>\*

**\* This is a required element.**

Contacts are people (or positions, like Data Manager) that should be contacted for access to, or information about, the data package. For Jornada data packages this should include:

* The “Current responsible investigator” (if available)
* The JRN LTER Information Manager (datamanager.jrn.lter@gmail.com)

JRN LTER defines the “Current responsible investigator” as the LTER principal investigator who curates a study and its associated data package(s). For data packages without a “current responsible PI”, John Anderson has historically been listed with this role under <contact>. When updating one of these packages, ask John if he would like to remain in this position. If not, he may be removed and it is acceptable to list only the Data Manager.

No other contacts are defined.

### <creator>\*

**\*This is a required element.**

Creators are people with direct intellectual contributions to the data package and so could include PIs, postdocs, students, and other researchers. There can be multiple <creator> elements in a Jornada EML document. These should include:

* Original PIs
* Former responsible PIs
* Current responsible PIs
* PostDocs, grad students, and other researchers (??? need to verify this)

If for some reason the original PI for the data package is unknown, it is acceptable to list the Jornada Basin LTER as an organization in the <creator> element.

### <project>

This element should be included in EML for all JRN data packages and should contain the <title>, <organization>, <funding>, and <personnel> sub-elements within it. List the Jornada Basin LTER as the project <title>, New Mexico State University as <organization>, and the current responsible PI for the data package as the <personnel>. The <funding> element should include the NSF award number (DEB-1832194 or similar) that was active when the project was initiated.

This element is created automatically by EMLassemblyline if a personnel entry is listed with the role “PI”.

*Provisional change*: If there is no current responsible PI for a completed package, it may be acceptable to omit the <project><personnel> element as long as the other elements are complete. (??? need to verify this is ok)

### Other personnel and organization elements

Other personnel and organization elements, such as <metadataProvider>, <associatedParty>, or <publisher>, can be defined in EML and have been included in Jornada Basin LTER data packages. There is no standardized policy for these yet, so when updating such packages, preserve these elements if possible.

## <pubDate>

Default for this at EDI and LTER sites appears to be that <pubDate> refers to the date of the latest metadata revision for a data package posted on EDI.

At JRN, this element has had other meanings (first date of publication), but there isn’t currently a consistent policy.

## <abstract>

A descriptive abstract of the data package, with enough context for a user to decide whether it is useful to them, should be included here. **Keep in mind that package abstracts are used in full-text searches**

Best practices:

* Include a brief statement about whether data collection is “ongoing” or “completed”.
* Include a brief description of data collection and data package update frequency (“updated monthly/yearly/every spring”).
* Include information on whether the data were collected on the “Jornada Experimental Range”, on the “CDRCC”, or on both.

## <keywordSet>

Multiple <keywordSet> elements can be defined. Optionally they may be labeled with a specific vocabulary they are taken from using the <keywordThesaurus> tag. The Jornada Basin LTER uses these keyword thesauri:

* [LTER Controlled Vocabulary](http://vocab.lternet.edu/vocab/vocab/index.php) - required to describe the subject matter of each data package
* [Jornada-specific thesauri](https://github.com/jornada-im/jrn_metadata_standards) - currently there are 3, which are still in development.
	+ “JRN Dataset Keywords” include subject matter, funding, time span, and other JRN LTER specific terms.
	+ “Jornada Project Names” include the descriptive names of abbreviations of studies or projects taking place in Jornada Basin (affiliated with JRN LTER or not)
	+ “Jornada Place Names” include the place names of important study locations or geographic features. These overlap with the “Jornada Place Names” in many cases.
* [LTER Core Areas](https://lternet.edu/core-research-areas/)
* EML creators can also define keywords independent of these vocabularies (no <keywordThesaurus> tag).

From the Jornada-specific thesauri, be sure to include:

* At least one subject matter term from the “JRN Dataset Keyword” list.
* Any relevant terms from the “Jornada Project Names” list
* Any relevant placenames from the “Jornada Place Names” list
* The Jornada study ID as “study {NNN}”, where {NNN} is the 3 digit study ID within the package ID (middle 3 digits). List this as a “Jornada Project Name” keyword.

The keywords “core” and “signature” may be added to EML documents, but the plan at this point is to keep a separate list of packages with these designations and then cross-reference those lists during searches of the JRN data catalog.

## <intellectualRights>

The Jornada Basin LTER uses the CC-BY attribution license, appended with the LTER network’s [suggested data access policy text](https://lternet.edu/data-access-policy/). There is a template in EMLassemblyline for this.

## <coverage>

A <coverage> element can be defined at multiple levels in the EML document: for a <dataset> element (/eml:eml/dataset/coverage), for a data entity (eml:eml/dataset/[entity]/coverage), within a <methods> element (eml:eml/dataset/methods/sampling/studyExtent/coverage), etc. Within each <coverage> element, three types of coverage elements can be defined - <geographicCoverage>, <taxonomicCoverage>, and <temporalCoverage>.

For now, at JRN LTER, we are focusing on providing adequate <coverage> metadata at the <dataset> element level, and there aren’t known standards for elements below this level (???). This is described in the sections below.

### <geographicCoverage>

Coordinates for Jornada data packages are obfuscated (deliberately, due to security concerns) in most cases. How, and how much, to provide spatial data is a matter of debate at JRN LTER. The general practice is to provide one or more bounding boxes surrounding the research sites in a data package, but these bounding boxes should not be detailed enough to allow casual visitors to EDI to identify sensitive research sites or instrumentation. Some packages, however, already have finer details in the <geographicCoverage> elements. In these cases the person packaging the data should probably preserve this metadata, and has some discretion on how to proceed after consulting with PIs. The general guideline for this element is outlined below.

For <dataset> elements, define at least one <geographicCoverage> element. This <geographicCoverage> element should contain at least one bounding box that encloses all research locations for all data entities in the package. When possible, multiple bounding boxes may be provided as additional <geographicCoverage> elements that describe other study or sampling areas (these may be associated with locational variables in a data entity if you are clever about it). The bounding boxes should be large enough to prevent casual users from identifying the sites (on the EDI map interface, for example).

Each <geographicCoverage> element should have a <geographicDescription> element associated that describes the location defined by the bounding boxes (or points, as the case may be) and include the text “Higher resolution spatial data for this data package can be obtained by contacting the Data Manager”.

### <taxonomicCoverage>

There are many systems for denoting taxonomic entities or groups, including local systems that are specific to the Jornada Basin LTER or its researchers. JRN data packages that contain taxonomic data (data attributes of species, groups of species, or other taxa) should provide ways to identify the taxonomic entities or groups in a <dataset> to modern, accepted taxonomic classifications at the *Genus species* or finer levels.

These links may be provided in a variable or category in the data entity itself (a “Genus\_species” variable, or example), as long as they are properly attributed to an authority like ITIS or USDA PLANTS. Local (JRN specific) taxonomic keys or coding systems that appear in a data entity must be available to the data user. In these cases the keys or coding systems used should be described in the metadata and provided as either an additional data entity in the data package, or as a link to other JRN EDI packages with this information.

Some current Jornada taxonomic keys are:

1. [Key to vascular plants in the Jornada Basin](https://portal.lternet.edu/nis/metadataviewer?packageid=knb-lter-jrn.210520001.1) (an EDI package)

### <temporalCoverage>

Provide the start and end date of the period over which the data were collected. For <dataset> elements this should include the data in all data entities (all .csv or other data files) that the EML document defines.

## <maintenance>

Describes whether data collection is ongoing or completed (in <description> element) and may include descriptions of the frequency of data and metadata changes (in <frequency> element).

1. Describe the data package as “completed” if data collection has ended.
2. Describe the data package as “ongoing” if data collection continues.
3. For “ongoing” data packages, a description of how often package revisions, including data and metadata changes, occur should be added to the <frequency> element under <maintenance>. For example, if new data are QA/QC’d, appended to a data entity, and published for a data package every month, the <frequency> element should describe these “monthly” revisions.
4. Optionally, and in addition to #3 above, the <frequency> element under <maintenance> can describe the frequency (hourly, daily, bimonthly, etc.) of the data in any data entities in the package. This may be especially important tabular time series.

## <methods>

This element should include a detailed description of how the data were collected or otherwise derived (field procedures, laboratory analysis, data synthesis and analysis). It should be concise but sufficient to reproduce the resulting data entity in the package.

Many Jornada packages have <methods> elements that refer to ancillary procedures documents, QA/QC specifications, data reduction scripts, etc. Whenever possible these should be preserved and archived as additional data entities in the data package (or at least a link to another repository should be provided).

If needed <methods> elements can be defined for individual data entities (/eml:eml/dataset/[entity]/methods). A hierarchy of additional elements, such as <instrument>, <sampling>, <methodStep>, <qualityControl>, can be also defined within each <methods> element. The GCE Toolbox+ does some of this for us with Jornada met data. In general though, descriptive text with optional pointers to other methods documentation is usually adequate.

## Data entities

These define the actual data files in the package (.csv files, rasters, etc). There can be more than one in a package.

Several possible elements are used to define individual data entities within a <dataset>. These include <dataTable>, <spatialRaster>, <spatialVector>, <storedProcedure>, <view>, and <otherEntity>. Each of these must contain other elements that describe the data they contain (the EntityGroup tree). The most important of these, for us, is the <attributeList> that defines the variables in a <dataTable> entity, since most Jornada data is tabular.

There are not any Jornada-specific metadata standards for data entities (???maybe there should be) or the elements they contain (<attributeList>). Use EML best practices.

# Information management, metadata, and EML resources

The LTER network (funded by NSF) maintains a [Data Access policy](https://lternet.edu/data-access-policy/) that forms the basis for how Information Managment (IM) systems should function at sites in the LTER network. LTER maintains an [IM web portal](https://im.lternet.edu) with some resources on IM practices, guidelines, and training materials. The [Guidelines for LTER IM Systems](https://im.lternet.edu/im_requirements/ims_guidelines) document encapsulates the guidelines and best practices for what an LTER IM system should be.

Recently, many of the resources on curating and publishing data and metadata packages has been pushed to [EDI](https://environmentaldatainitiative.org). EDI’s “5 Phases of data publishing” is a large collection of resources, but in particular, [Phase 3](https://environmentaldatainitiative.org/resources/five-phases-of-data-publishing/phase-3/) provides useful information on EML metadata best practices like the [EML Best Practices V3 document](https://environmentaldatainitiative.files.wordpress.com/2017/11/emlbestpractices-v3.pdf).

The EML specification is defined and maintained by the [Knowledge Network for Biocomplexity](https://knb.ecoinformatics.org). They provide the EML Schema and additional (human readable) information here:

<https://knb.ecoinformatics.org/external//emlparser/docs/index.html>