Jornada Metadata Standards for EML Creation

The Information Management Team

27 June 2019

Table of Contents

[Introduction 2](#_Toc12549372)

[Metadata standards by EML element 2](#_Toc12549373)

[title 2](#_Toc12549374)

[Personnel and organization elements (creator, contact, etc.) 2](#_Toc12549375)

[contact 2](#_Toc12549376)

[creator 3](#_Toc12549377)

[Other personnel and organization elements 3](#_Toc12549378)

[pubDate 3](#_Toc12549379)

[abstract 3](#_Toc12549380)

[keywordSet 3](#_Toc12549381)

[intellectualRights 4](#_Toc12549382)

[coverage 4](#_Toc12549383)

[geographicCoverage 4](#_Toc12549384)

[taxonomicCoverage 4](#_Toc12549385)

[temporalCoverage 4](#_Toc12549386)

[maintenance 4](#_Toc12549387)

[methods 5](#_Toc12549388)

[project 5](#_Toc12549389)

[Data entities 5](#_Toc12549390)

[Information management, metadata, and EML resources 5](#_Toc12549391)

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

# Introduction

This document defines metadata standards to use when creating 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.

Note that EML documents are heirarchical. The top level element is eml:eml, and within that are contained 3 elements: access, dataset, and additional\_metadata. The metadata standards in this document refer to the dataset element of your EML file only. Maybe we’ll add standards on access and additional\_metadata at a later time.

# Metadata standards by EML element

## title

Descriptive title that includes type of data collected, geographic location, and time range of the data (what, where, when). In general we use “Jornada” in some form as part of the geographic description. It may also be useful to include a project name or abbreviation if the package is part of a collection from a large project (NEAT, SMES, etc).

## Personnel and organization elements (creator, contact, 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 entity must be supplied in every EML document. All EML documents should usually have at least one creator entity, but exceptions can be made if this information has been lost in very old Jornada datasets.

### 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 as a contact with this role. When updating one of these packages, ask John if he would like to remain in this position. If not he may be removedand it is acceptable to list only the Data Manager.

No other contacts are defined.

### creator

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 elements in a Jornda EML document. These should include:

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

### 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 EML documents. There is no standardized policy for these yet, so when updating such packages, preserve these elements if possible.

## pubDate

No JRN policy yet.

## 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**

## keywordSet

Multiple keywordSet elements can be defined. Optionally they may be labeled with the specific vocabulary they are identified 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) - still in development
* EML creators can also define keywords independent of these vocabularies (no keywordThesaurus tag.

## 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: the dataset level (/eml:eml/dataset/coverage), data entity level (eml:eml/dataset/[entity]/coverage), within the methods (eml:eml/dataset/methods/sampling/studyExtent/coverage), etc. Within each coverage element, three types of coverage are possible - geographicCoverage, taxonomicCoverage, and temporalCoverage.

For now, at JRN LTER, we are focusing on providing adequate coverage metadata for the full dataset, not really at finer details. That is what is described below

### geographicCoverage

Coordinates for Jornada data packages are obfuscated (deliberately, due to security concerns) in most cases.

At the dataset level define at least one bounding box that includes all research locations for all data entities in the package. When possible, multiple bounding boxes may be provided to describe multiple study or sampling areas (these may be associated with locational variables in a data entity if you are clever about it).

The geographicDescription attached to this bounding box should describe the location defined by the bounding boxes and include the text “Higher resolution spatial data for this data package can be obtained by contacting the Data Manager”.

### taxonomicCoverage

No JRN policy yet.

### temporalCoverage

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

## maintenance

1. Describe the data package as “completed” if data collection has ended.
2. Describe the data package as “ongoing” if data collection continues.
3. If “ongoing”, add the frequency of data collection for the package (hourly, daily, bimonthly, etc.)

## methods

This is a required element that 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 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 heirarchy 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.

## project

Some Jornada data packages have this element, but it is not required and there are no standards for it (that I am aware of).

## 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 in a dataset. These include dataTable, spatialRaster, spatialVector, storedProcedure, view, and otherEntitymay. They 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.

I’m not aware of any Jornada-specific metadata standards for data entities. 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>