Network Information System

Documenting Source and Derived Data Products using EML for the Trends Data Warehouse

Request for Comments - Version 1.0

M. Servilla - servilla@lternet.edu

LTER Network Office
Department of Biology, MSC03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

Introduction

Use of the Ecological Metadata Language (EML) for storing metadata from the LTER Trends Network Information System data module has been assumed from the project's inception. A specific plan on how EML would be used, however, was not addressed until now. The following white paper describes an approach to storing metadata in the EML for both derived and source data products as defined the Trends project. It concludes with a contrived example that documents the metadata for both the derived product and originating source products in a single EML document. This is a "living" paper and will evolve through the planning stage of the Trends Phase II project.

Model Perspective

From a technical perspective, the Trends information system is designed around a data warehouse (Figure 1). Data in the warehouse consists of two types: (1) ***source data*** – a data product that is obtained from the originating site; and (2) ***derived data*** – a data product resulting from some type of transformation applied to the source product. Note that the transformation process may not actually change or alter the source data; it is, however, still considered "derived". A derived data product may consist of one or more source data products. Source data are maintained in the primary database. Derived data are maintained in the secondary database. Data transformation occurs between the primary and secondary databases.

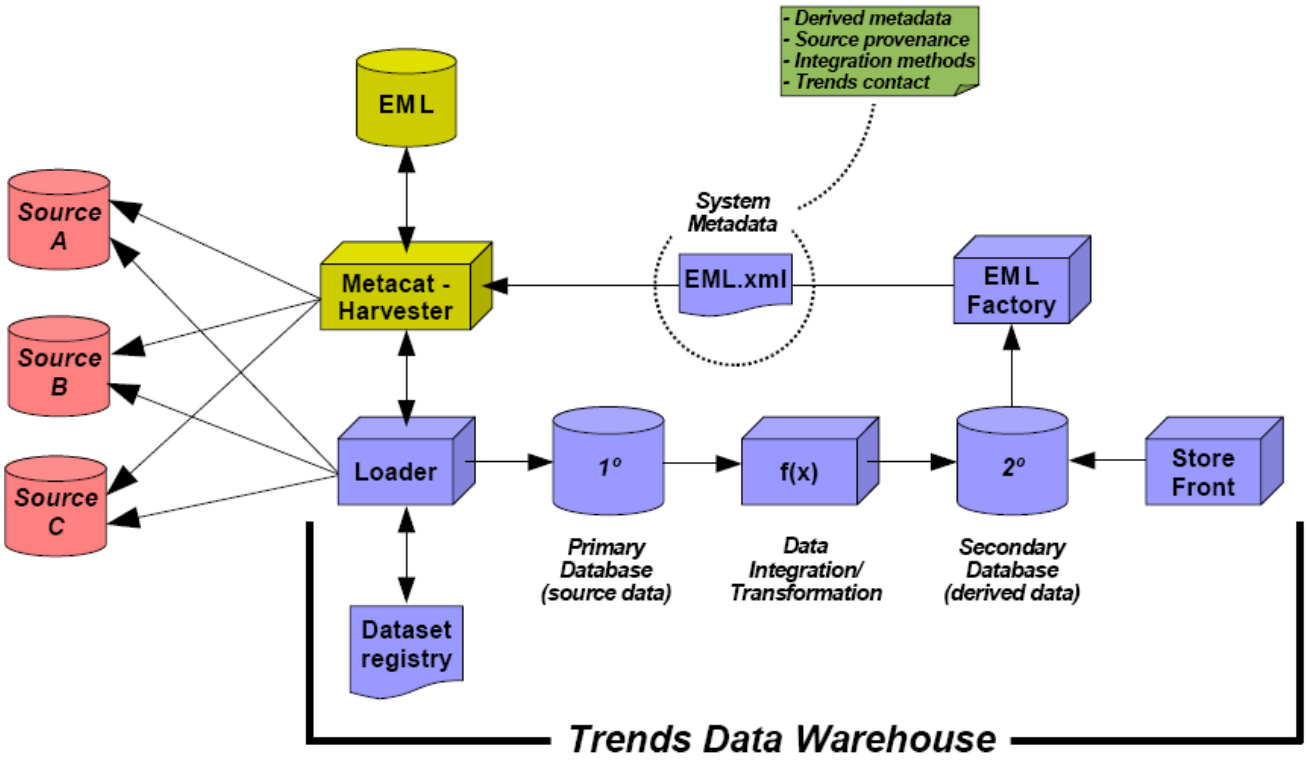

Figure 1 - Conceptual view of the Trends Data Warehouse.

It is assumed that source metadata is stored in an EML document within Metacat. Metadata describing the derived product is stored in both the secondary database and also as an

EML document that is stored in Metacat. The EML document that describes the derived product consists of three significant components: (1) the dataset description for the metadata of the derived product (***MDP***), (2) the dataset description for metadata of all source products (***MSP***), and (3) a description of any (if applicable) transformation process (***FX***) applied to the source data to create the derived data. Within the EML document, the MDP component is the principal or highest level *dataset* subtree of the schema (Figure 2). The MDP component contains information pertaining only to the derived product that is found in the secondary database. Specific information found in the *dataset* subtree, such as *creator* or *contact* is only related to the

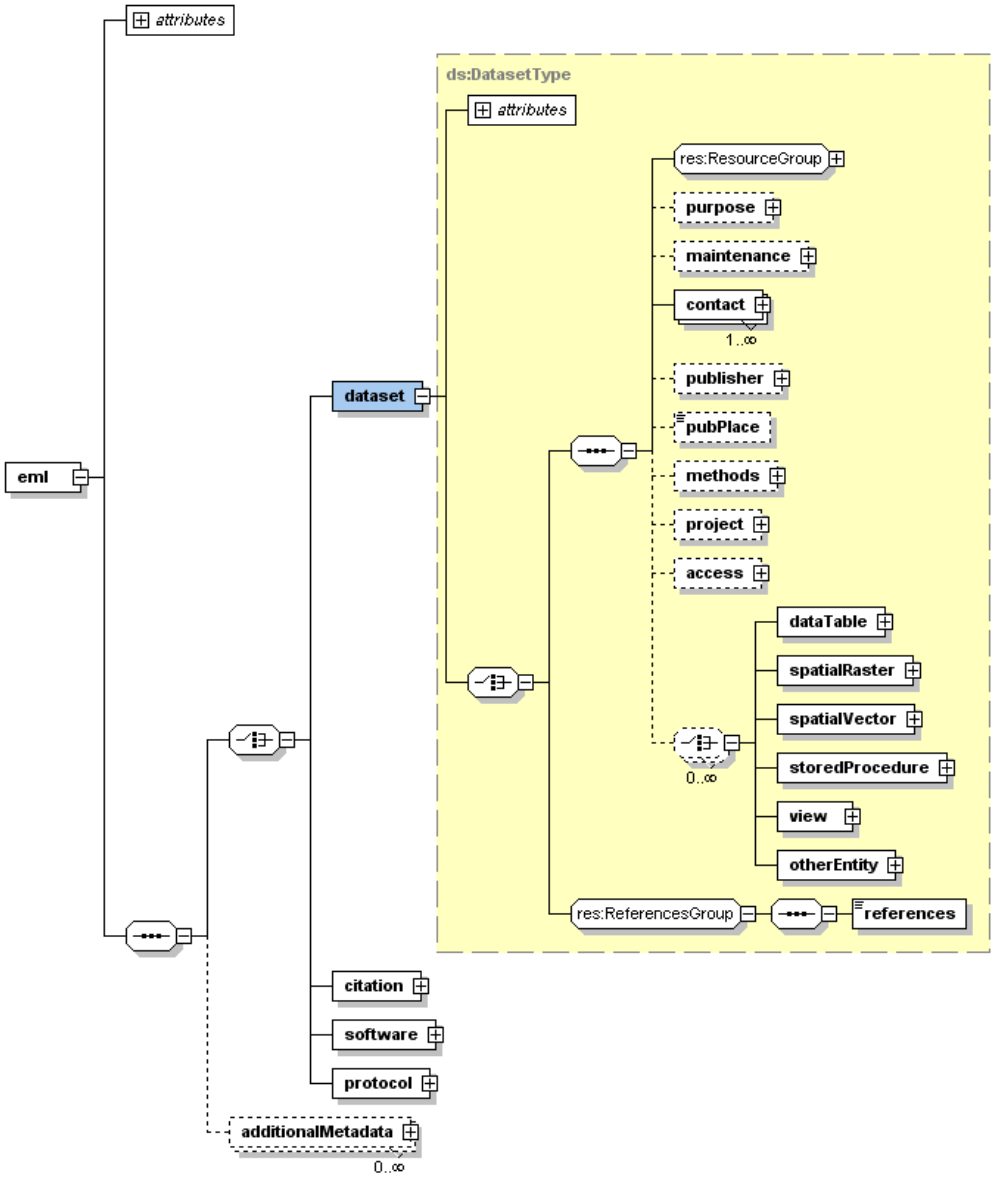

Figure 2 - The EML *dataset* subtree.

Trends Data Warehouse – there should not be any reference to the original source product other than perhaps a textual description within the *abstract*.

The *methodStep* subtree (Figure 3) has two key roles in the EML document – to document all MSP and to document any FX. The MSP component is effectively managed within the *methodStep* subtree by providing a high-level textual description of the source product in the *description* (this is a required element) and by defining the appropriate fields in the *dataSource*. The *dataSource* element is a declared as a *dataset* type. This means that a minimal set of information from the source product EML document must be copied to all of the required elements within the *dataset* subtree (these include: *title*, *creator*, and *contact*). Most important, however, is that the remainder of the source product EML document can be referenced as an online URL through the element series *distribution*::*online*::*url* within the resource group. The URL content would simply reference the source product EML document contained in Metacat (e.g., http://metacat.lternet.edu/knb/metacat?action=read&qformat=xml&docid=knb-lter-vcr.20.2).

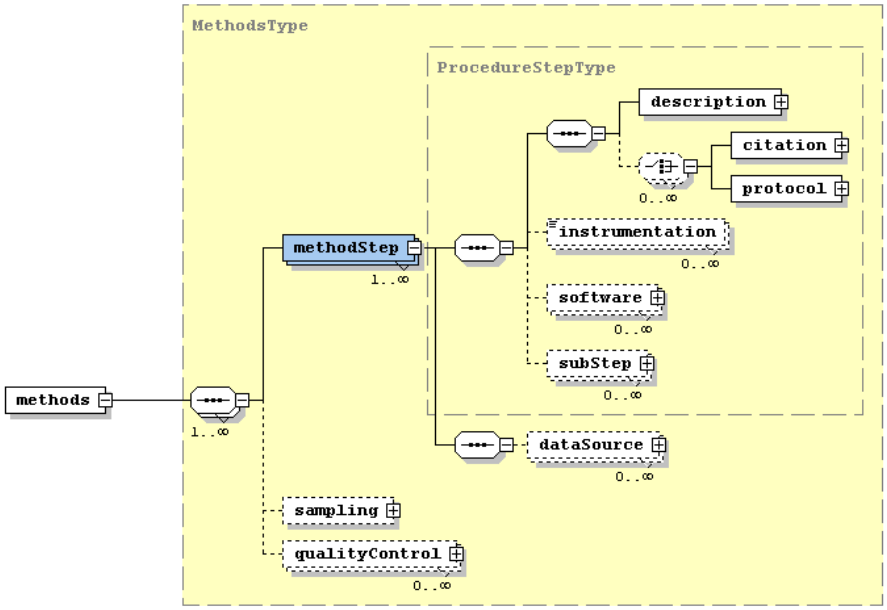

Figure 3 - The EML *methodStep* subtree.

Example

Now, let's address this issue with a simple fictitious example. Let us assume that we are creating a derived product for Net Primary Productivity (NPP) that is calculated based on the amount of dry Carbon measured in one year. Our derived product will use $kg/m^2/year$ for its unit. We will create our derived product by combining data from several research sites. Unfortunately, each site uses a slightly different unit formula, which requires us to perform a simple transformation on the source data products so that they will be the same as our derived product unit. The source sites and products are described below, in Table 1.

Site Name	*Source Unit*	*Conversion to Derived Unit*
ONL	kg/m^2/year	1
NRJ	g/m^2/year	1/1000
VES	kg/cm^2/year	1/10000
SAECN	kg/m^2/month	12

Table 1 - Unit definitions for each site and their associated conversion factor.

We assume that each research site generates an EML document that describes their unique method of determining NPP, and their EML document is harvested into the Metacat when ever new or modified data is available. We collect their 100 year data individually (we could have automated this phase since their EML document tells us where to access the source data), place it into an "R" data frame, and apply the appropriate conversion factors. Now we have our derived data product. Further, we now have to generate our EML document that succinctly describes our derived data product, including information pertaining to the source data products and conversion process that we applied to the source data. Because we are the creator of the derived data product, we will take the lead role in the EML document and put our information in the primary *dataset* subtree. To document the source data, we place the appropriate information into a *methods* subtree, one subtree for each of the source data products. Finally, we also use the methods subtree to document our conversion software – in this case, an "R" script. The complete EML document (Figure 4) is then harvested into the Metacat along with all of the other EML documents.

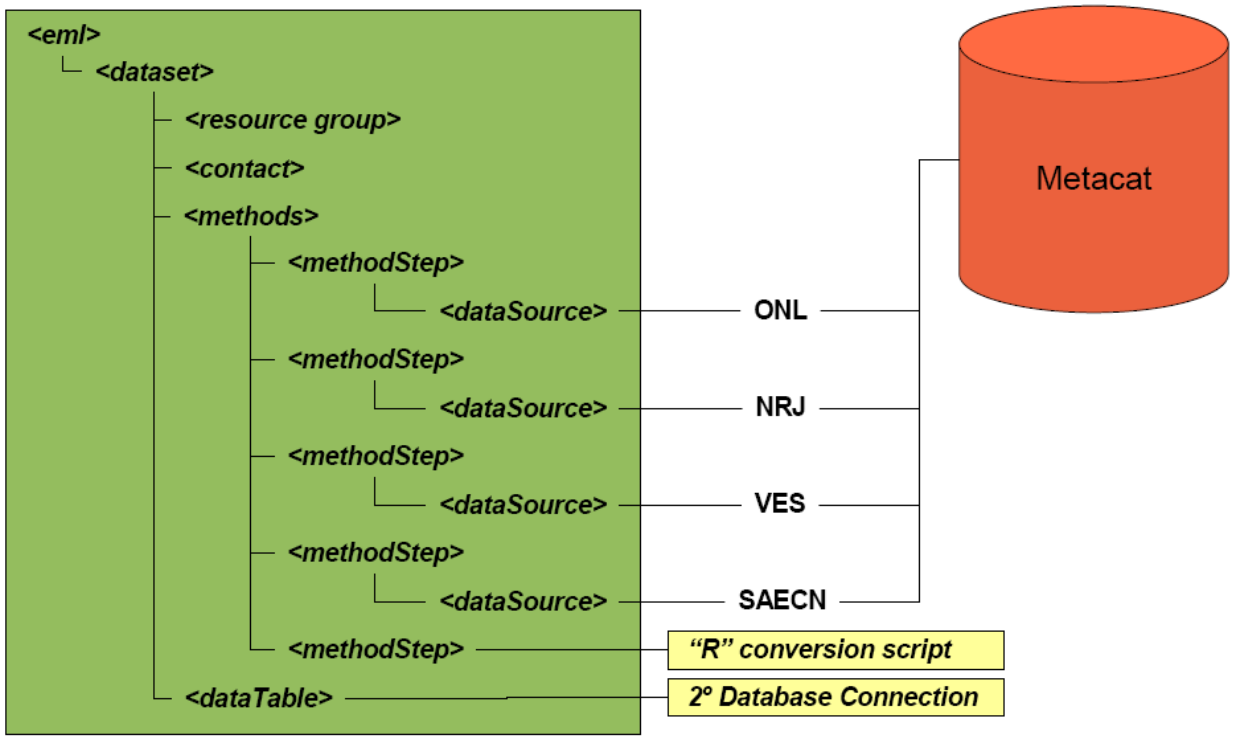

Figure 4 - Generalized EML structure for derived data product.

Conclusion

There are a number of benefits to the approach outlined above. First, the cardinality of the *methodStep* subtree is 1..∞, therefore it can nicely capture one or more MSPs and one or more FXs. Second, the MDP should reference only the MSPs that were used to create the derived product. Since the the URL uses the specific revision number of the MSP EML document, the MDP will always reference the correct MSP EML document (even if the MSP EML document has been deprecated by a new revision). As such, both metadata for the derived data product and all of the source data products can be captured in a single EML document.

Appendix 1 – Example EML document

```
<?xml version="1.0" encoding="UTF-8"?>
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.0.1"
xmlns:ds="eml://ecoinformatics.org/dataset-2.0.1"
xmlns:doc="eml://ecoinformatics.org/documentation-2.0.1"
xmlns:cit="eml://ecoinformatics.org/literature-2.0.1"
xmlns:prot="eml://ecoinformatics.org/protocol-2.0.1"
xmlns:res="eml://ecoinformatics.org/resource-2.0.1"
xmlns:sw="eml://ecoinformatics.org/software-2.0.1"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="eml://ecoinformatics.org/eml-2.0.1
C:\eml-2.0.1\eml.xsd" packageId="knb-trends.23.1" system="knb" scope="system">
      <dataset>
            <title>
                  Derived NPP data from ONL, NRJ, VES, and SAECN
                  field stations from 1906 through 2006.
            </title>
            <creator>
                  <organizationName>
                        LTER Trends Institute, LLC.
                  </organizationName>
                  <address>
                        <deliveryPoint>Box 30003</deliveryPoint>
                        <deliveryPoint>MSC 3JER</deliveryPoint>
                        <deliveryPoint>NMSU</deliveryPoint>
                        <city>Las Cruces</city>
                        <administrativeArea>NM</administrativeArea>
                        <postalCode>88003-0003</postalCode>
                        <country>USA</country>
                  </address>
                  <electronicMailAddress>
                        trendsdev@lternet.edu
                  </electronicMailAddress>
            </creator>
            <abstract>
                  <section>
                        <para>
                              The following synthesis process
                              generates a derived NPP product for
                              four Long-Term Ecological Research
                              sites by converting site source specific
                              values of NPP to the global schema
                              defined for the Trends NPP derived product.
                        </para>
                  </section>
            </abstract>
            <contact>
                  <positionName>Trends Information Manager</positionName>
                  <organizationName>
                        LTER Trends Institute, LLC.
                  </organizationName>
                  <address>
                        <deliveryPoint>Box 30003</deliveryPoint>
                        <deliveryPoint>MSC 3JER</deliveryPoint>
                        <deliveryPoint>NMSU</deliveryPoint>
                        <city>Las Cruces</city>
                        <administrativeArea>NM</administrativeArea>
                        <postalCode>88003-0003</postalCode>
```

```
            <country>USA</country>
      </address>
      <electronicMailAddress>
            trends@lternet.edu
      </electronicMailAddress>
</contact>
<methods>
      <methodStep>
            <description>
                  <section>
                        <para>
                              We utilize NPP data collected
                              from 1906 to 2006 from the ONL
                              LTER site. The ONL NPP data unit
                              definition is kg/m^2/yr. This
                              unit does not require conversion.
                        </para>
                  </section>
            </description>
            <dataSource>
                  <title>
                        NPP data from ONL 1906 to 2006
                  </title>
                  <creator>
                        <organizationName>
                              ONL LTER
                        </organizationName>
                  </creator>
                  <distribution>
                        <online>
                              <url>
                                    http://metacat.lternet.edu
                                    /knb/metacat?action=read&amp;
                                    qformat=xml&amp;docid=
                                    knb-lter-onl.23.1
                              </url>
                        </online>
                  </distribution>
                  <contact>
                        <organizationName>
                              ONL LTER
                        </organizationName>
                        <positionName>
                              ONL Information Manager
                        </positionName>
                        <electronicMailAddress>
                              im@onl.lternet.edu
                        </electronicMailAddress>
                  </contact>
            </dataSource>
      </methodStep>
      <!-- similar methodStep subtrees here for the
           remaining three sites: NRJ, VES, and SAECN -->
      <methodStep>
            <description>
                  <section>
                        <para>
                              The "R" programming environment
                              is used to convert source NPP data
                              to the same unit type used in the
                              derived NPP global schema.
```

```
                              </para>
                        </section>
                        <section>
                              <para>
                                    "R" script goes here
                              </para>
                        </section>
                  </description>
            </methodStep>
      </methods>
      <dataTable>
            <entityName>Net Primary Productivity</entityName>
            <attributeList>
                  <attribute>
                        <attributeName>DT</attributeName>
                        <attributeDefinition>
                              Date and time stamp in UTC format
                        </attributeDefinition>
                        <measurementScale>
                              <datetime>
                                    <formatString>
                                          YYY-MM-DDThh:mm:ss
                                    </formatString>
                              </datetime>
                        </measurementScale>
                  </attribute>
                  <attribute>
                        <attributeName>NPP</attributeName>
                        <attributeDefinition>
                              Net Primary Productivity
                        </attributeDefinition>
                        <measurementScale>
                              <ratio>
                                    <unit>
                                          <standardUnit>
                                                kilogramsPerMeter
                                                SquaredPerYear
                                          </standardUnit>
                                    </unit>
                                    <numericDomain>
                                          <numberType>
                                                real
                                          </numberType>
                                    </numericDomain>
                              </ratio>
                        </measurementScale>
                  </attribute>
            </attributeList>
      </dataTable>
  </dataset>
</eml:eml>
```