Understanding the influence a community recommendation has on an organization’s metadata

Sean Gordon, Ted Habermann The HDF Group

Table of Contents

[Highlights 1](#_Toc476038474)

[Abstract 1](#_Toc476038475)

[Abbreviations 1](#_Toc476038476)

[Keywords 2](#_Toc476038477)

[Introduction 2](#_Toc476038478)

[Metadata Standards/Dialects/Recommendations 2](#_Toc476038479)

[Recommendations and Dialects at DataONE 3](#_Toc476038480)

[LTER Recommendations 3](#_Toc476038481)

[Data 3](#_Toc476038482)

[DataOne Member Nodes 3](#_Toc476038483)

[Process 3](#_Toc476038484)

[Process 6](#_Toc476038485)

[Results 7](#_Toc476038486)

## **Highlights**

* Comparison of EML usage across DataONE
* Metadata recommendations can be used to improve the contents of a record

## **Abstract**

Many organizations make use of structured documentation that is machine-readable. This metadata makes discovery, access, use, and understanding much easier to write after the entire article has been written.

## **Abbreviations**

* EML, Ecological Metadata Language;
* LTER, Long Term Ecological Network;
* KNB;
* XML;
* XSLT;
* CSDGM, Content Standard for Digital Geographic Metadata;

## **Keywords**

* LTER network;
* Metadata completeness;
* Ecological metadata language;
* Information management;
* DataONE;
* Collection analysis;
* Community recommendations;
* Metadata dialects;
* Data Analysis

# Introduction

All scientists and scientific communities recognize the need to document observations and processing clearly and completely to support understanding and reproducibility of their scientific results. Many datasets and products are documented using approaches and tools developed by data collectors to support their own analysis and understanding needs. This documentation can exist almost any conceivable form, each with associated storage and preservation strategies. This custom, often unstructured, approach may work well for independent investigators or in the confines of a laboratory or community, but it makes it difficult for users outside of these small groups to discover, use, and understand the data without consulting with its creators.

Metadata, in contrast to documentation, provides well‐defined content in structured representations. This makes it possible for users to access and quickly understand many aspects of datasets that they need to answer specific questions, but have not collected or created themselves. It also makes it possible to integrate information into discovery and analysis tools, and to provide consistent references from the metadata to external documentation.

## Metadata Standards/Dialects/Recommendations

Scientific communities that recognize the need for metadata typically address that need using one of a couple approaches: they either use a metadata standard proposed by a related community or organization, or they develop one that fits their needs. These metadata standards include concept names, definitions and associated structures. A concept is a general term for describing a documentation entity, typically an element of attribute in XML. In most cases, they also include a standard representation for the metadata. We refer to these representations as *metadata dialects*. Typically, the communities or organizations that develop the standard also develop a set of recommendations for metadata content. We refer to these as *metadata recommendations*.

Organizations also create recommendations for dialects they did not create but have implemented in their community. The Long Range Ecological Network created the LTER Recommendation for Completeness to help guide the creation of Ecological Metadata Language records.

There are five levels in the LTER recommendation: Identification, Discovery, Evaluation, Access, and Integration. All levels of the LTER recommendation are subsets of concepts in the EML dialect. LTER uses the EML dialect (D1) and created a recommendation with 5 levels (R1, R2, R3, R4, R5) Four concepts from the Identification level (R1) are EML schema required concepts: Resource Title, Resource Identifier, Author / Originator, and Resource Contact. (R6) A second community creates a dialect (D2) with recommendations at 2 levels (R7, R8). As the overlap between these dialects and recommendations show, common documentation needs exist, particularly for the discovery use case.



## Recommendations and Dialects at DataONE

The DataONE Metadata Repository (REFERENCE) provides a unique opportunity to explore relationships between metadata recommendations and dialects. It includes collections of metadata records from XX different Member Nodes in XX different dialects. The most common dialects are EML and CSDGM, which is commonly known as FGDC because the U.S. Federal Geographic Data Committee developed the standard.

Ecological Metadata Language was developed by the Knowledge Network for Biocomplexity (KNB) and the Long-Term Ecological Research (LTER) Program (REFERENCE) to address specific needs of the ecological research community. The authors were influenced by both FGDC and ISO metadata standards, so EML shares characteristics of both of these standards. Many ecological research groups in the U.S. and around the world actively use EML.

As the ecological research community gained experience with EML, it became clear that, in many cases, metadata records were not complete or consistent enough to serve important community requirements. In order to address this problem, a group of LTER metadata experts developed a set of recommendations for metadata content (REFERENCE). These recommendations included elements expected to cover five important use cases: xxx.

The LTER recommendations were well publicized and supported in the LTER community, so we might expect that the LTER metadata records are more complete with respect to these recommendations than other EML collections. The DataOne Repository includes many EML collections and thus provides an excellent test case for understanding the impact of recommendations across communities. We might expect that LTER metadata requirements overlap many other DataOne member node communities and, therefore, that the LTER metadata recommendations would be relevant for many DataOne member nodes. This is the hypothesis we explore in this paper.

## LTER Recommendations

# Data

## DataOne Member Nodes

## Process

Complete, consistent, and usable documentation is critical for facilitating discovery and reuse of scientific data, particularly if you use a metadata dialect that has been standardized. There are many metadata recommendations from organizations like the OGC, FGDC, NASA, and LTER, that can provide documentation guidance. Often, the recommendations that organizations develop are for a specific metadata dialect.

Table 1 - A dialect is a community specific instantiation of the documentation language that is specific to a community.

|  |
| --- |
| Known Metadata Dialects in DataONE |
| DataCite 3.1 |
| Content Standard for Digital Geospatial Metadata (CSDGM) |
| Content Standard for Digital Geospatial Metadata (CSDGM) Biological Data Profile |
| ISO 19115 and ISO 19115-2 / ISO 19139 and ISO 19139-2 |
| Dryad |
| OneDCX |
| Mercury Metadata Standard |
| Attribute Convention for Data Discovery (ACDD) |
| Ecological Metadata Language (EML) |

The concepts being described are similar, and are often the same between dialects. We consider a concept a general term for describing a documentation entity. For example, many different dialects use a XML element to refer to the resource’s title in their shared documentation record. Since there are many such concept similarities, we can quantitatively report on a collection in many dialects for many recommendations using the concepts contained in the collection’s records.

Table 2 - A recommendation is a set of concepts that an organization considers to be important in documenting scientific discovery.

|  |  |
| --- | --- |
| **Metadata Recommendations** | **Originating** **Organization** |
| [CSW\_Discovery](http://wiki.esipfed.org/index.php/Data_Discovery_(CSW)) | Open Geospatial Consortium |
| [ISO-1\_Discovery](http://wiki.esipfed.org/index.php/Data_Discovery_(ISO-19115-1)) | International Standards Organization |
| [DIF\_Discovery](http://wiki.esipfed.org/index.php/Data_Discovery_(DIF)) | National Aeronautics and Space Administration |
| FGDC\_Discovery | Federal Geographic Data Committee |
| DataCite\_Discovery | DataCite |
| DCAT\_Discovery | World Wide Web Consortium |
| ECS\_Discovery | National Aeronautics and Space Administration |
| ECHO\_Discovery | National Aeronautics and Space Administration |
| ACDD\_Discovery | University Corporation for Atmospheric Research / ESIP Documentation Cluster |
| LTER\_Completeness | Long Term Ecological Research Network |
| Dryad-Package | Dryad Digital Repository |

The HDF Group and NCEAS use the metadata in the DataONE repository to research the effect that use of metadata recommendations have on a collection’s metadata quality as part of the MetaDIG project. In MetaDIG each of the dialects used by DataONE Member Nodes are the collections. We use recommendation completeness as a quantitative measure of a collection’s quality according to the recommendation’s originating organization.

Table 3 - A collection is a group of metadata records, commonly organized by a data center, organization or project and often stored in a database or web accessible folder.

|  |  |
| --- | --- |
| MemberNode | Dialects Used |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

In the first phase of our research, we analyzed the Member Nodes that had EML and CSDGM dialect collections using the LTER Recommendation. We created a conceptual version of the LTER recommendation at a high level. Instead of looking for the creator’s email address and organization we test only for the creator. We used the main concepts present in the five levels of the LTER Recommendation to assess the collections for completeness of documentation. The five levels are Identification, Discovery, Evaluation, Access, and Integration. As you can see in the chart below, EML contains every concept in each of these levels while CSDGM is missing one concept in each level except for Access. comparing across the Member Nodes using CSDGM or EML.

## Process

We created a sample of up to 250 records from each member node, and separated these by dialect version. After cleaning up the resultant collections a report was generated on each. These reports detailed each xpath that contained a text value. The reports were concatenated by dialect and fed into an Excel workbook. The workbook allowed us to calculate the average occurrence count of each element, as well as collection level average occurrence for a dialect.

By selecting the elements in the five levels we were able to compare completeness across member nodes. Below the Discovery level is shown. Notice that it is a only counting complete concepts. These concepts are ones that have an average occurrence count of 1 or higher

## Results

Process