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

Sean Gordon, Ted Habermann The HDF Group

Table of Contents

[Highlights 1](#_Toc476310554)

[Abstract 1](#_Toc476310555)

[Abbreviations 1](#_Toc476310556)

[Keywords 2](#_Toc476310557)

[Introduction 2](#_Toc476310558)

[Metadata Standards/Dialects/Recommendations/Concepts 2](#_Toc476310559)

[Dialects and Recommendations at DataONE 2](#_Toc476310560)

[LTER Recommendation 3](#_Toc476310561)

[Comparison of DataONE dialects and recommendations 4](#_Toc476310562)

[Data 5](#_Toc476310563)

[DataONE Member Nodes 5](#_Toc476310564)

[Methods 6](#_Toc476310565)

[Process 7](#_Toc476310566)

[Results 8](#_Toc476310567)

[Concept Occurrence Percentages 8](#_Toc476310568)

[Dialect Maximum compared with Member Node completeness 11](#_Toc476310569)

[Signature Score Groups by Dialect 12](#_Toc476310570)

[Incomplete Concepts by Dialect 13](#_Toc476310571)

[Complete Concept Counts 14](#_Toc476310572)

[Conclusions and Further Questions 17](#_Toc476310573)

[Bibliography 17](#_Toc476310574)

## Highlights

* Comparison of EML usage across DataONE
* Metadata recommendations 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;
* BDP, Biological Data Profile

## 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 that make it easier to share and discover. 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/Concepts

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*.

## Dialects and Recommendations at DataONE

The DataONE Data Catalog (“DataONE Data Catalog,” n.d.) provides a unique opportunity to explore relationships between metadata recommendations and dialects. It includes collections of metadata records from 26 different Member Nodes in 8 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 (“The Long Term Ecological Research Network | Long-term, broad-scale research to understand our world,” n.d.)

to address specific needs of the ecological research community. The authors were influenced by both FGDC and ISO metadata standards, so EML shares characteristics with both 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. To address this problem, a group of LTER metadata experts developed a set of recommendations for metadata content (EML Best Practices, n.d.). These recommendations included elements expected to cover five important use cases: Identification, Discovery, Evaluation, Access, and Integration.

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 Recommendation

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. Each of the levels recommend specific concepts designed to provide information about the dataset for a specific use case. All levels of the LTER recommendation are subsets of concepts in the EML dialect. This is illustrated in Diagram 1. 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.



Table 0 - Conceptual description of the recommendations

|  |  |  |
| --- | --- | --- |
| Recommendation Level | # Concepts | Concept Titles |
| Identification | 11 | Resource Identifier, Resource Title, Author / Originator, Metadata Contact, Contributor Name, Publisher, Publication Date, Resource Contact, Abstract, Keyword, Resource Distribution |
| Discovery | 4 | Spatial Extent, Taxonomic Extent, Temporal Extent, Maintenance |
| Evaluation | 5 | Resource Use Constraints, Process Step, Project Description, Entity Type Definition, Attribute Definition |
| Access | 2 | Resource Access Constraints, Resource Format |
| Integration | 3 | Attribute List, Attribute Constraints, Resource Quality Description |

## Comparison of DataONE dialects and recommendations

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.

# Data

## DataONE Member Nodes

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 DIBBs project. In the DIBBs MetaDIG project 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 1 - A dialect is a community specific instantiation of the documentation language that is specific to a community.

|  |
| --- |
| 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 |
| Ecological Metadata Language (EML) |

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 | # Records | Collection Dialects and Counts |
| CDL | 250 | CSDGM(250) |
| CLOEBIRD | 1 | EML2.1.0(1) |
| DRYAD | 251 | Dryad(251) |
| EDACGSTORE | 250 | CSDGM(250) |
| EDORA | 28 | Mercury(28) |
| ESA | 53 | EML2.1.1(5), EML2.0.1(17), EML2.1.0(31) |
| GLEON | 13 | EML2.1.1(12)EML2.0.1(1) |
| GOA | 98 | EML2.1.1(98) |
| IARC | 250 | DataONE Dublin Core Extended v1.0(250) |
| IOE | 24 | EML2.1.1(24) |
| KNB | 250 | EML2.1.1(31) EML\_Access\_module\_version\_2.0.0beta6(15), EML \_Dataset\_module\_version\_2.0.0beta4(2), EML \_Dataset\_module\_version\_2.0.0beta6(13), EML \_Physical\_module\_version\_2.0.0beta6(2), EML2.0.0(101), EML2.0.1(49), EML2.1.0(35) |
| KUBI | 172 | EML2.1.1(172) |
| LTER | 250 | EML2.0.1(18), EML2.1.0(146), EML2.1.1(86) |
| LTER\_EUROPE | 165 | EML2.1.1(165) |
| NMEPSCOR | 7 | CSDGM(7) |
| ONEShare | 109 | EML2.1.1(109) |
| ORNLDAAC | 250 | Mercury(250) |
| PISCO | 250 | EML2.0.1(250) |
| RGD | 248 | Mercury(248) |
| SANPARKS | 247 | EML2.0.0(9), EML2.0.1(16), EML2.1.0(222) |
| SEAD | 18 | CSDGM(18) |
| TERN | 250 | EML2.1.1(250) |
| TFRI | 251 | EML2.1.1(17), EML2.1.0(28), EML2.0.1(206), |
| USANPN | 6 | EML2.1.1(6) |
| USGSCSAS | 250 | CSDGM(240), BDP(10) |
| US\_MPC | 250 | DataONE Dublin Core Extended v1.0(250) |

## Methods

Crosswalks Workflow

Step Zero: Follow the instructions on the following pages to prepare your computer.

* [How to Add Saxon Home Edition](https://github.com/scgordon/CrosswalksWorkflow/wiki/How-to-Add-Saxon-Home-Edition)
* [How To Install Anaconda](https://github.com/scgordon/CrosswalksWorkflow/wiki/How-to-Install-Anaconda)

Step One: Create Branches for the Analysis

* Create branches in both repositories entitled "organizationName\_recTag" eg: DataOne\_LTER or NASA\_UMM-C. This will create a snapshot of all of the data that goes into an analysis.
* Extract dialects used at data center, compare resultant definitions with concepts the recommendation contains to prepare AllCrosswalks.xml to create accurate rubrics.
* Step Two: AllCrosswalks.xml Improvements
* Enhancing and creating dialects using the [Dialect Extractor](https://github.com/tedhabermann/Crosswalks/wiki/Dialect-Extractor) and [Merge Dialect](https://github.com/tedhabermann/Crosswalks/wiki/Merge-Dialect) tools.
* Ingesting and testing of concept references in a recommendation component. Since the set of concepts that are contained in the different parts of the recommendation are the most important concepts for the analysis, compare a list of references in the recommendation

Step Three: Rubric Creation

* Create rubrics using the [crosswalks](https://github.com/tedhabermann/Crosswalks/wiki/Creating-Rubrics) tool
* This can be facilitated through the use of the batch script [writeAllRubricTransforms.sh](https://github.com/scgordon/CrosswalksWorkflow/wiki/writeAllRubricTransforms.sh) which runs specific instances of [rubricTransform.sh](https://github.com/scgordon/CrosswalksWorkflow/wiki/rubricTransform.sh)
* Test rubrics

Step Four: Ingest Metadata

* Ingest the organization's collections into the directory "rawMetadata". Then copy these collection sub-directories into a sub-directory of CrosswalksWorkflow/collections e.g. "CrosswalksWorkflow/collections/NASA" and begin to clean the directory structure and ensure namespace agreement for the prefix used in AllCrosswalks.

Step Five: Create json

* Using the batch script, [jsonCreator.sh](https://github.com/scgordon/CrosswalksWorkflow/wiki/jsonCreator.sh), leverage [runTransform.sh](https://github.com/scgordon/CrosswalksWorkflow/wiki/runTransform.sh) to create the json needed for your analysis.

Step Six: Create xlsx

* Use the [generateSS.sh](https://github.com/scgordon/CrosswalksWorkflow/wiki/Using-generateSS.sh) script to run the python scripts needed to combine the json. The resultant file is deposited into Reports under the recommendation tag.

Step Seven: Create Visualizations

## 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.

In the first phase of our research, we analyzed the Member Nodes that had EML 2.1.1 and CSDGM dialect collections using the LTER Recommendation . We used the Mercury and BDP collections as CSDGM. We created a conceptual version of the LTER recommendation at a high level detailed in Table 0. 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.

# Results

At a high level, LTER has more complete metadata than other member nodes in DataONE that use EML 2.1.1 or various CSDGM versions. This completeness is characterized as a concept occurrence percentage for each of the member nodes. There is a comparison to the recommendation’s dialect maximum, identification of signature scores, and Incomplete concept counts. These visualizations are intended to give an overview of the collections completeness with respect to the first three levels of the LTER recommendation for the sample set that was downloaded from the DataONE Data Catalog.

## Concept Occurrence Percentages

Concept occurrence tables show what percentage of the collection’s records contain the dialect definition for that concept.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| EML | ESA | GLEON | GOA | IOE | KNB | LTER | LTER\_EUROPE | ONEShare | TERN | TFRI | USANPN | KUBI | DataOne |
| Resource Identifier | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Resource Title | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Author / Originator | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Metadata Contact | 100% | 58% | 0% | 0% | 68% | 98% | 84% | 0% | 0% | 0% | 0% | 0% | 34% |
| Contributor Name | 100% | 42% | 95% | 0% | 74% | 1% | 0% | 0% | 0% | 53% | 100% | 0% | 39% |
| Publisher | 0% | 25% | 0% | 0% | 0% | 100% | 0% | 94% | 100% | 0% | 0% | 0% | 27% |
| Publication Date | 100% | 50% | 0% | 0% | 68% | 100% | 69% | 100% | 0% | 0% | 0% | 0% | 41% |
| Resource Contact | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Abstract | 100% | 92% | 100% | 100% | 97% | 100% | 88% | 98% | 100% | 100% | 100% | 0% | 90% |
| Keyword | 80% | 75% | 100% | 96% | 87% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 95% |
| Resource Distribution | 100% | 100% | 97% | 100% | 87% | 100% | 100% | 94% | 100% | 100% | 100% | 100% | 98% |
| Taxonomic Extent | 100% | 0% | 77% | 8% | 35% | 0% | 21% | 0% | 100% | 12% | 0% | 0% | 29% |
| Spatial Extent | 100% | 92% | 94% | 100% | 90% | 100% | 48% | 97% | 100% | 100% | 100% | 100% | 93% |
| Temporal Extent | 100% | 92% | 94% | 4% | 87% | 100% | 98% | 94% | 100% | 35% | 100% | 100% | 84% |
| Maintenance | 0% | 25% | 0% | 0% | 0% | 99% | 0% | 0% | 0% | 0% | 0% | 0% | 10% |
| Resource Use Constraints | 100% | 92% | 100% | 100% | 94% | 99% | 89% | 88% | 0% | 82% | 100% | 0% | 79% |
| Process Step | 80% | 67% | 94% | 0% | 68% | 100% | 100% | 0% | 100% | 88% | 100% | 0% | 66% |
| Project Description | 0% | 33% | 95% | 8% | 13% | 1% | 0% | 94% | 100% | 0% | 0% | 0% | 29% |
| Entity Type Definition | 0% | 75% | 79% | 8% | 16% | 2% | 0% | 95% | 0% | 24% | 100% | 0% | 33% |
| Attribute Definition | 0% | 83% | 84% | 29% | 23% | 3% | 0% | 95% | 0% | 100% | 100% | 0% | 43% |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CSDGM | CDL | USGSCSAS | EDACGSTORE | EDORA | ORNLDAAC | RGD | SEAD | NMEPSCOR | DataOne |
| Resource Identifier | -100% | -100% | -100% | -100% | -100% | -100% | -100% | -100% | -100% |
| Resource Title | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Author / Originator | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Metadata Contact | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Contributor Name | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Publisher | 100% | 26% | 1% | 0% | 0% | 0% | 67% | 0% | 24% |
| Publication Date | 100% | 100% | 100% | 0% | 0% | 0% | 100% | 100% | 63% |
| Resource Contact | 100% | 80% | 100% | 100% | 100% | 100% | 67% | 100% | 93% |
| Abstract | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Keyword | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Resource Distribution | 0% | 100% | 100% | 100% | 100% | 100% | 67% | 100% | 83% |
| Taxonomic Extent | -100% | -100% | -100% | -100% | -100% | -100% | -100% | -100% | -100% |
| Spatial Extent | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Temporal Extent | 0% | 36% | 95% | 100% | 100% | 100% | 89% | 57% | 72% |
| Maintenance | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Resource Use Constraints | 100% | 100% | 100% | 0% | 0% | 0% | 100% | 100% | 63% |
| Process Step | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| Project Description | -100% | -100% | -100% | -100% | -100% | -100% | -100% | -100% | -100% |
| Entity Type Definition | 100% | 98% | 81% | 0% | 0% | 0% | 0% | 100% | 47% |
| Attribute Definition | 100% | 98% | 81% | 0% | 0% | 0% | 0% | 100% | 47% |

## Dialect Maximum compared with Member Node completeness

## Signature Score Groups by Dialect

## Incomplete Concepts by Dialect

## Complete Concept Counts

# Conclusions and Further Questions

# Bibliography

DataONE [WWW Document], n.d. URL https://www.dataone.org/ (accessed 3.1.17).

DataONE Data Catalog [WWW Document], n.d. URL https://search.dataone.org/#data (accessed 3.1.17).

The Long Term Ecological Research Network | Long-term, broad-scale research to understand our world [WWW Document], n.d. URL https://lternet.edu/ (accessed 3.1.17).