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

Sean Gordon, Ted Habermann The HDF Group

Table of Contents

[Highlights 2](#_Toc478726987)

[Abstract 2](#_Toc478726988)

[Abbreviations 2](#_Toc478726989)

[Keywords 3](#_Toc478726990)

[Introduction 3](#_Toc478726991)

[Metadata Standards/Dialects/Recommendations/Concepts 3](#_Toc478726992)

[Dialects and Recommendations at DataONE 4](#_Toc478726993)

[LTER Recommendation 4](#_Toc478726994)

[Comparison of DataONE dialects and the LTER Recommendation 6](#_Toc478726995)

[Data 6](#_Toc478726996)

[Dialects 7](#_Toc478726997)

[DataONE Member Node Sampling 7](#_Toc478726998)

[Methods 8](#_Toc478726999)

[Process 9](#_Toc478727000)

[Results 10](#_Toc478727001)

[Concept Occurrence Percentages 11](#_Toc478727002)

[Identification Level 11](#_Toc478727003)

[Discovery Level 12](#_Toc478727004)

[Evaluation Level 13](#_Toc478727005)

[Access Level 13](#_Toc478727006)

[Integration Level 14](#_Toc478727007)

[Level Completeness by Collection 15](#_Toc478727008)

[Identification Level 15](#_Toc478727009)

[Discovery Level 16](#_Toc478727010)

[Evaluation Level 17](#_Toc478727011)

[Access Level 18](#_Toc478727012)

[Integration Level 19](#_Toc478727013)

[Signature Scores 20](#_Toc478727014)

[Average Signature Score Sums 20](#_Toc478727015)

[Signature Sum Distribution 21](#_Toc478727016)

[Signature Scores by Recommendation Level 22](#_Toc478727017)

[Identification Level 23](#_Toc478727018)

[Discovery Level 23](#_Toc478727019)

[Evaluation Level 24](#_Toc478727020)

[Access Level 25](#_Toc478727021)

[Integration Level 26](#_Toc478727022)

[Conclusions and Further Questions 27](#_Toc478727023)

[Observation 1 27](#_Toc478727024)

[Conclusion 1 27](#_Toc478727025)

[Observation 2 27](#_Toc478727026)

[Conclusion 2 27](#_Toc478727027)

[Observation 3 27](#_Toc478727028)

[Conclusion 3 27](#_Toc478727029)

[Observation 4 27](#_Toc478727030)

[Conclusion 4 27](#_Toc478727031)

[Observation 5 27](#_Toc478727032)

[Conclusion 5 28](#_Toc478727033)

[Questions 28](#_Toc478727034)

[Bibliography 28](#_Toc478727035)

## Highlights

* Collections in EML and CSDGM measured by a conceptual version of the LTER Recommendation for Completeness
* Comparison of EML and CSDGM usage across DataONE
* Metadata recommendations as a community activity to improve completeness
* Quantitative measures of recommendation completeness

## Abstract

Many organizations make use of structured documentation that is machine-readable. Metadata makes discovery, access, use, and understanding of scientific datasets possible. Organizations and communities have created recommendations for metadata. These recommendations are often dialect specific. By rewriting the recommendations conceptually, quantitative analysis of the structures of multiple dialects becomes possible. This is a study of the LTER recommendation for Completeness and CSDGM and EML records in DataONE. The purpose of the study is to determine if LTER created more complete metadata by utilizing a community recommendation for documentation of datasets through providing guidance and tools for the metadata creators.

## Abbreviations

* EML, Ecological Metadata Language;
* LTER, Long-Term Ecological Research Network;
* KNB, Knowledge Network for Biocomplexity ;
* CLOEBIRD, ;
* ESA, ;
* GLEON, ;
* GOA, ;
* IOE, ;
* KUBI, ;
* LTER\_Europe, ;
* ONEShare, ;
* PISCO, ;
* SANPARKS, ;
* TERN, ;
* TFRI, ;
* USANPN, ;
* OneDCX,  DataONE Dublin Core Extended v1.0;
* XML, ;
* XSLT, ;
* CSDGM, Content Standard for Digital Geographic Metadata;

## Keywords

* LTER network;
* Metadata completeness;
* Ecological metadata language;
* Content Standard for Digital Geographic Metadata;
* Information management;
* DataONE;
* Collection analysis;
* Community recommendations;
* Metadata dialects;
* Data Analysis;
* Concept Occurrence;
* Collection Coverage
* Collection Convergence

# 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. In most cases, they also include a standard representation for the metadata. We refer to these representations as *metadata dialects*. These metadata dialects include concept names, definitions and associated structures. A *concept* is a general term for describing a documentation entity, typically a defined element or attribute in XML. 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 6 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.

EML was developed by KNB and LTER (“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 collections in DataONE. The DataONE Repository includes many EML and CSDGM 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 metadata 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, or need. 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 as defined by USGEO. In LTER these concepts are found in the Identification and Discovery levels.



The conceptual design of the recommendation allows records in other dialects to be analyzed by the same recommendation. The blue concept names are concepts that appear in the FGDC recommendation as well. 10 of the 25 concepts in the LTER recommendation are present in the FGDC recommendation. Many other concepts are closely related to concepts in the FGDC recommendation, such as Keyword and Spatial Extent. FGDC calls for Theme Keyword and Bounding Box.

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 the LTER Recommendation

Each level of the LTER recommendation contains metadata concepts needed for that documentation use case. 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. This means that a record at the CSDGM dialect maximum will never contain all the concepts in any of the levels except for access. CSDGM records can only be complete with respect to the CSDGM dialect maximum. CSDGM records will never be complete because there are concepts the dialect doesn’t contain. The *dialect maximum* is the number of concepts from a recommendation that a dialect contains. For example Mercury and BDP are other dialects in DataONE that extend CSDGM to contain taxonomic information in the case of BDP, or an identifier for the resource in Mercury’s case. In these cases, organizations have extended CSDGM when it did not contain the concepts they needed to describe in their metadata. The dialect maximum for BDP in the Discovery level of the LTER Recommendation is the same as the *recommendation maximum*, or count of concepts in a recommendation level.

# Data

The HDF Group and NCEAS use the metadata in the DataONE repository to research the effect that use of a metadata recommendation by a community have on a collection’s metadata completeness as part of the DIBBs project. We use recommendation completeness as a quantitative measure of a collection’s quality according to the recommendation’s originating organization. By comparing record collections from other parts of the greater community with the LTER collection sample Since there are no quality measures of the contents of the record, records may show as incomplete, even though they contain all the relevant information for that dataset. Perhaps the most common example of a concept like this is Taxonomic Extent. Taxonomic Extent may not be needed for a project because nothing biological is being measured. DataONE has many dialects and member nodes. Before describing the results, here is a description of the data and the methods.

## Dialects

In the DIBBs MetaDIG project each of the dialect versions used by DataONE member nodes are separated into collections. The following table contains the abbreviation and name of the dialects in the DataONE sample set of metadata. Dialects are often referred to as a metadata language. By using dialect to describe these standards, the similarities rather than the distinctions are highlighted.

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

|  |
| --- |
| Metadata Dialects in the DataONE Sample |
| Content Standard for Digital Geographic Metadata (CSDGM) |
| Biological Data Profile of CSDGM (BDP) |
| Dryad Metadata Schema, (Dryad) |
| DataONE Dublin Core Extended v1.0 (OneDCX) |
| Mercury Metadata Standard (Mercury) |
| Ecological Metadata Language (EML) |

## DataONE Member Node Sampling

DataONE is adding member nodes to the repository. These nodes contained metadata in October of 2015 when the sample was taken. The following table describes the record counts received from the sampling of the DataONE repository, as well as what dialect version the documents are written in. The record count for each member node is the total of all the different dialects and dialect versions described in the Dialect Collections and Counts column.

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.

|  |  |  |
| --- | --- | --- |
| Member Node | Records | Dialect Version Collections 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 | OneDCX (250) |
| IOE | 24 | EML2.1.1 (24) |
| KNB | 250 | 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),  EML2.1.1 (31) |
| 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 | 248 | EML2.0.1 (248) |
| 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 | 250 | EML2.1.1 (17),  EML2.1.0 (27),  EML2.0.1 (206), |
| USANPN | 6 | EML2.1.1 (6) |
| USGSCSAS | 250 | CSDGM (240),  BDP (10) |
| US\_MPC | 250 | OneDCX (250) |

## Methods

Crosswalks Workflow is a step-by-step process that is used to analyze the meta-dataset for completeness, using several recommendations and dialects. It is described in detail in the Crosswalks Workflow GitHub repository’s wiki pages (Gordon, 2016). Some steps require permission as access to files in private GitHub repositories is required. A brief explanation follows.

The first step is to define the dialect and the recommendation conceptually. This prepares the system for testing the collection with the recommendation. Once the recommendation and the dialect are defined, the metadata records get organized into a directory structure.

When records are shared via xml that are close to standard but have some simple differences, the tools will return a rubric showing no concepts contained in the records. It can be simple to clean these records so that the tools will locate the concepts, providing a better analysis. Sometimes records will have a namespace prefix added that is not part of the dialect or will be empty files. Since EML uses the same prefix for all versions, sometimes the version needs to be altered in the files so they all match up. This is done so that the rubric created for the dialect recommendation pair can read and score the records in the collection accurately.

The rubric, which is an xsl transform reads the records individually and creates a json scorecard that is then fed into a python script that creates the spreadsheet. This spreadsheet is then used to create the visualizations used to explain the completeness of the collections and the comparison between them.

## Process

We created a sample of up to 250 records from each member node at DataONE. Collections were separated by dialect version and member node. This was done using a python script, created at NCEAS (Mecum, 2015).

We created a conceptual version of the LTER recommendation at a high level detailed in Table 0. We used the main concepts present in the five levels of the LTER Recommendation to assess the collections for completeness of documentation. We used the EML 2.1.1 schema (reference) to identify EML dialect definitions for the HDF concept ontology.

A decision was made to utilize the records from all the different EML versions except the beta versions at KNB. The beta versions do not share a root with standard EML. The collections were combined into a single directory for each member node. The namespace prefix “eml” was modified to the EML 2.1.1 version in each record written in a previous version. The collections were then treated as though they were EML 2.1.1 as the LTER recommendation had been in use through all the different versions found in the sample set. The resultant collections, record counts, and collection dialects are described in the following table.

Table 4 – Collections ready for analysis

|  |  |  |
| --- | --- | --- |
| Member Node | Record Count | Dialect |
| CLOEBIRD | 1 | EML |
| ESA | 53 | EML |
| GLEON | 13 | EML |
| GOA | 98 | EML |
| IOE | 24 | EML |
| KNB | 218 | EML |
| KUBI | 172 | EML |
| LTER | 250 | EML |
| LTER\_EUROPE | 165 | EML |
| TERN | 250 | EML |
| TFRI | 250 | EML |
| USANPN | 6 | EML |
| ONEShare | 109 | EML |
| PISCO | 248 | EML |
| SANPARKS | 247 | EML |
| SEAD | 18 | CSDGM |
| EDACGSTORE | 250 | CSDGM |
| CDL | 250 | CSDGM |
| NMEPSCOR | 7 | CSDGM |
| USGSCSAS | 240 | CSDGM |

After cleaning up the resultant collections a json report was generated on each record. These reports detailed the presence or absence of the concept’s dialect definition. The reports were concatenated by collection and fed into a python script in a private repository. The script creates an Excel workbook that details the presence/absence and count of each concept in the LTER recommendation for each record. The workbook allowed us to calculate the average occurrence count of each element, as well as collection level average occurrence for a dialect. Visualizations are created using this data. By identifying the records that contain the concepts in the five levels we compare completeness across member nodes in DataONE that use CSDGM and EML to measure if and how LTER used their recommendation to improve the community’s metadata completeness.

# Results

At a high level, the LTER organization’s EML sample does not appear to be uniformly more complete than other member nodes in DataONE. However, it also indicates that LTER has the most complete records. Recommendation completeness for a collection is characterized as a concept occurrence percentage for each of the member nodes. Recommendation coverage can also be observed from the concept occurrence tables. There is an identification of signature score groups and a distribution of LTER records throughout the signature score group. Recommendation completeness can also be seen at the record level, as a sum of the recommendation level signature score. Finally, the average completeness of a concept in each of the 5 recommendation levels are displayed for each collection. These visualizations are intended to measure recommendation completeness with respect to the LTER recommendation for the CSDGM and EML records in 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. The visualization is comprised of rows for each recommendation concept and columns for each dialect. Cells are filled with a color or a percentage. The percentage is how many records in the sample set contain that concept.

Green means every record in the member node’s collection contains the concept. Yellow represents 0%, a concept that the dialect contains but is not in any record in the member node’s collection. Red represents a concept that is not contained in any record in the collection. Furthermore, the concept cannot be documented within the structure of the collection’s current dialect. The tables are intended to show how complete a collection is for a recommendation level.

### Identification Level

The identification level of the LTER recommendation is primarily complete for all member nodes, regardless of documentation dialect. There are incomplete concepts in each of the member nodes collections. Each member node has at least one concept from the level that is unused or unusable in the dialect the collection is documented in, except LTER. The LTER member node collection contains at least one record where each of the concepts in the level occurs. Even the CSDGM records have a high occurrence percentage for schema required concepts: Resource Title, Resource Identifier, Author / Originator, and Resource Contact.

From a qualitative perspective, concept occurrence percentages for a collection with 250 records, like LTER, may need to be weighted more than a collection with 1 record that has the same concept occurrence percentage. This study utilizes quantitative measures for completeness, so questions of quality cannot be easily addressed.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Identification | Resource Identifier | Resource Title | Author / Originator | Metadata Contact | Contributor Name | Publisher | Publication Date | Resource Contact | Abstract | Keyword | Resource Distribution |
| CLOEBIRD | 100% | 100% | 100% | 100% | 0% | 100% | 100% | 100% | 100% | 100% | 0% |
| ESA | 100% | 100% | 100% | 100% | 94% | 0% | 100% | 100% | 100% | 94% | 100% |
| GLEON | 100% | 100% | 100% | 54% | 46% | 23% | 46% | 100% | 92% | 77% | 62% |
| GOA | 100% | 100% | 100% | 0% | 95% | 0% | 0% | 100% | 100% | 100% | 0% |
| IOE | 100% | 100% | 100% | 0% | 0% | 0% | 0% | 100% | 100% | 96% | 0% |
| KNB | 100% | 100% | 100% | 56% | 53% | 1% | 18% | 100% | 94% | 89% | 56% |
| KUBI | 100% | 100% | 100% | 0% | 0% | 0% | 0% | 100% | 0% | 100% | 0% |
| LTER | 100% | 100% | 100% | 83% | 18% | 86% | 94% | 100% | 99.2% | 99% | 36% |
| LTER\_EUROPE | 100% | 100% | 100% | 84% | 0% | 0% | 69% | 100% | 88% | 100% | 100% |
| ONEShare | 100% | 100% | 100% | 0% | 0% | 94% | 100% | 100% | 98% | 100% | 94% |
| PISCO | 100% | 100% | 100% | 0% | 91% | 0% | 0% | 100% | 100% | 100% | 99% |
| SANPARKS | 100% | 100% | 100% | 2% | 32% | 0% | 2% | 100% | 85% | 97% | 2% |
| TERN | 100% | 100% | 100% | 0% | 0% | 100% | 0% | 100% | 100% | 100% | 100% |
| TFRI | 100% | 100% | 100% | 0% | 31% | 0% | 0% | 100% | 99% | 99% | 0% |
| USANPN | 100% | 100% | 100% | 0% | 100% | 0% | 0% | 100% | 100% | 100% | 0% |
| CDL | -100% | 100% | 100% | 100% | 0% | 100% | 100% | 100% | 100% | 100% | 0% |
| EDACGSTORE | -100% | 100% | 100% | 100% | 7% | 1% | 100% | 100% | 100% | 100% | 100% |
| NMEPSCOR | -100% | 100% | 100% | 100% | 100% | 0% | 100% | 100% | 100% | 100% | 100% |
| SEAD | -100% | 100% | 100% | 100% | 50% | 67% | 100% | 67% | 100% | 100% | 67% |
| USGSCSAS | -100% | 100% | 100% | 100% | 42% | 24% | 100% | 79% | 100% | 100% | 100% |

### Discovery Level

In the discovery level, there are only 4 concepts. Only one of these concepts is used every collection, Spatial Extent. Temporal Extent is only missing in CDL, but most collections have an incomplete usage of the concept, whereas with Spatial Extent over half of the collections contain Spatial Extent in every record. Just under half of the collections don’t use Taxonomic Extent at all, and every CSDGM record does not contain taxonomic information, as the dialect doesn’t have the structural location for the concept. Most records do not have Maintenance information. Except for 3 records from GLEON and one from CLOEBIRD, the 138 records from LTER are the only records not written in CSDGM that have Maintenance information. CSDGM records all contain the Maintenance concept.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Discovery | Spatial Extent | Taxonomic Extent | Temporal Extent | Maintenance |
| CLOEBIRD | 100% | 100% | 100% | 100% |
| ESA | 92% | 70% | 100% | 0% |
| GLEON | 92% | 0% | 92% | 23% |
| GOA | 94% | 77% | 94% | 0% |
| IOE | 100% | 8% | 4% | 0% |
| KNB | 92% | 23% | 86% | 0% |
| KUBI | 100% | 0% | 100% | 0% |
| LTER | 97% | 4% | 98% | 55% |
| LTER\_EUROPE | 48% | 21% | 98% | 0% |
| ONEShare | 97% | 0% | 94% | 0% |
| PISCO | 100% | 0% | 100% | 0% |
| SANPARKS | 98% | 15% | 95% | 0% |
| TERN | 100% | 100% | 100% | 0% |
| TFRI | 97% | 40% | 91% | 0% |
| USANPN | 100% | 0% | 100% | 0% |
| CDL | 100% | -100% | 0% | 100% |
| EDACGSTORE | 100% | -100% | 95% | 100% |
| NMEPSCOR | 100% | -100% | 57% | 100% |
| SEAD | 100% | -100% | 89% | 100% |
| USGSCSAS | 100% | -100% | 34% | 100% |

### Evaluation Level

The Evaluation level contains 5 concepts. Every collection except for KUBI use the Resource Use Constraints concept. The KUBI collection doesn’t contain any of the concepts in the Evaluation Level. Every CSDGM record is missing a project description. While individual records may describe the project as a larger work citation, not all larger work citations are project descriptions. Thus, the documentation dialect does not contain an explicit location for project description information. It is of note that 5 member nodes that use the EML dialect do not use the concept in their collections and only 4 collections exist where you can expect to see a project description at least 9 out of 10 times: GLEON, ONEShare, PISCO and TERN. The LTER sample only contains the concept in 40 records, or 16% of the sample.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation | Resource Use Constraints | Process Step | Project Description | Entity Type Definition | Attribute Definition |
| CLOEBIRD | 100% | 0% | 0% | 100% | 100% |
| ESA | 100% | 87% | 0% | 0% | 0% |
| GLEON | 92% | 69% | 38% | 69% | 85% |
| GOA | 100% | 94% | 95% | 79% | 84% |
| IOE | 100% | 0% | 8% | 8% | 29% |
| KNB | 95% | 62% | 11% | 13% | 20% |
| KUBI | 0% | 0% | 0% | 0% | 0% |
| LTER | 96% | 92% | 16% | 52% | 58% |
| LTER\_EUROPE | 89% | 100% | 0% | 0% | 0% |
| ONEShare | 94% | 0% | 94% | 95% | 95% |
| PISCO | 100% | 100% | 99% | 1% | 100% |
| SANPARKS | 44% | 57% | 2% | 13% | 69% |
| TERN | 100% | 100% | 100% | 0% | 0% |
| TFRI | 82% | 97% | 6% | 52% | 90% |
| USANPN | 100% | 100% | 0% | 100% | 100% |
| CDL | 100% | 0% | -100% | 100% | 100% |
| EDACGSTORE | 100% | 0% | -100% | 81% | 81% |
| NMEPSCOR | 100% | 0% | -100% | 100% | 100% |
| SEAD | 100% | 0% | -100% | 0% | 0% |
| USGSCSAS | 100% | 0% | -100% | 100% | 100% |

### Access Level

The Access level is close to complete for all of the collections documented in the CSDGM dialect. Only CDL and most of the SEAD collection are missing the Resource Format concept. LTER is close to complete in documenting constraints on accessing the resource but only 58% of records contain the resource’s format.

|  |  |  |
| --- | --- | --- |
| Access | Resource Access Constraints | Resource Format |
| CLOEBIRD | 100% | 100% |
| ESA | 68% | 0% |
| GLEON | 92% | 85% |
| GOA | 100% | 84% |
| IOE | 100% | 29% |
| KNB | 39% | 20% |
| KUBI | 0% | 0% |
| LTER | 93% | 58% |
| LTER\_EUROPE | 100% | 0% |
| ONEShare | 0% | 0% |
| PISCO | 0% | 100% |
| SANPARKS | 90% | 69% |
| TERN | 0% | 0% |
| TFRI | 18% | 90% |
| USANPN | 100% | 100% |
| CDL | 100% | 0% |
| EDACGSTORE | 100% | 100% |
| NMEPSCOR | 100% | 100% |
| SEAD | 100% | 6% |
| USGSCSAS | 100% | 100% |

### Integration Level

In the Integration level, there are only two member nodes that have collections containing every concept. LTER and KNB. Both of these collections are documented in the EML dialect. Both of these member nodes helped to create the EML dialect. No other member nodes even use the Attribute Constraints concept. TFRI is the only other EML using member node whose collection contains the Resource Quality Description Concept. All CSDGM collections contain the Resource Quality Description concept, but CSDGM does not document Attribute Constraints. Of the five collections that do not use the Attribute List concept, SEAD is the only member node that uses CSDGM.

|  |  |  |  |
| --- | --- | --- | --- |
| Integration | Attribute List | Attribute Constraints | Resource Quality Description |
| CLOEBIRD | 100% | 0% | 0% |
| ESA | 0% | 0% | 0% |
| GLEON | 85% | 0% | 0% |
| GOA | 84% | 0% | 0% |
| IOE | 29% | 0% | 0% |
| KNB | 20% | 1% | 1% |
| KUBI | 0% | 0% | 0% |
| LTER | 58% | 0.4% | 8% |
| LTER\_EUROPE | 0% | 0% | 0% |
| ONEShare | 95% | 0% | 0% |
| PISCO | 100% | 0% | 0% |
| SANPARKS | 69% | 0% | 0% |
| TERN | 0% | 0% | 0% |
| TFRI | 90% | 0% | 1% |
| USANPN | 100% | 0% | 0% |
| CDL | 100% | -100% | 100% |
| EDACGSTORE | 81% | -100% | 82% |
| NMEPSCOR | 100% | -100% | 100% |
| SEAD | 0% | -100% | 6% |
| USGSCSAS | 100% | -100% | 95% |

## Level Completeness by Collection

Level completeness percentage is calculated by taking the average of the level’s concept occurrence percentage for each concept. LTER is not the most complete in any level, but it is always more complete than the average for DataONE’s EML records. CSDGM records tend to be more complete than EML records, except in the Evaluation level. There are 10 concepts that are present in both the LTER and the FGDC recommendations as is detailed in Table 0.

### Identification Level

In the Identification Level, ESA has the most complete collection at 90%. LTER is next at 83%. NMEPSCOR and CLOEBIRD are 82% complete. Only 6 member nodes have less than two thirds completeness for the level.

### Discovery Level

Only 4 collections are more than two thirds complete. Two ofthese collections document in the EML dialect. CLOEBIRD’s collection of 1 record at 100% and TERN. EDACGSTORE and SEAD are the CSDGM collections. The CSDGM average is at 64% and is the same as LTER’s completeness. It is higher than the EML average of 54%.

### Evaluation Level

The Evaluation Level is the first level where a member node’s collection is missing every concept. KUBI does not use any of the concepts in the Evaluation level. GOA is the most complete member node at 90% complete for the level. No CSDGM documented collection is more than 60% complete. LTER and the EML average are more complete than the CSDGM average.

### Access Level

The Access level has 5 member nodes with 100% completeness. Two are EML collections and three are CSDGM collections.

### Integration Level

The Integration level is the least complete in both dialects. 4 collections do not contain any of the concepts.

## Signature Scores

These results are presented as counts of records with identical completeness scores with respect to the recommendation. The completeness scores are given in terms of the number of elements that are missing from a record, so low scores are good.

When a recommendation includes multiple levels (e.g. Mandatory, Recommended, and Optional), the scores are given as a series of numbers, one for each level. These are termed signatures. Typically, many records are missing the same concepts and, therefore, have identical signature score groups. Signature score groups are a way to expose the shining examples, the most complete metadata in the collection. They provide a good way to curate similarly documented metadata records and to chart a path towards completeness.

The signature 2 3 1 indicates a metadata record that has been tested for three levels and is missing 2 concepts from the first level, 3 from the second, and 1 concept from the third level. This record is less complete than a record with a signature of 1 1 1 and more complete than a record with a signature of 3 4 3.

## Average Signature Score Sums

One way to measure completeness is to take the signature scores and add the different levels up. This sum is the total missing concepts from the dialect maximum. In the case of an LTER signature score for a CSDGM record, the record can never be complete even if the signature score sum is 0, because the dialect maximum is 4 concepts less than the recommendation maximum. To normalize the signature score sums, the recommendation maximum is used. This means that where a CSDGM signature score sum used to be 0, it is not counted as 4 concepts missing.

Here are the average signature score sums. The list is ordered from most to least complete Average Signature Sum. Except for the one record from CLOEBIRD, LTER has the lowest average signature score sum of any EML collection. However, the most complete signature score sums come from member nodes that use the CSDGM dialect, except for the 1 CLOEBIRD record.

|  |  |  |  |
| --- | --- | --- | --- |
| Member Node | Dialect | Record Count | Average Signature Sum |
| CLOEBIRD | EML | 1 | 6.00 |
| NMEPSCOR | CSDGM | 7 | 6.43 |
| USGSCSAS | CSDGM | 240 | 7.28 |
| EDACGSTORE | CSDGM | 250 | 7.73 |
| LTER | EML | 250 | 7.99 |
| GOA | EML | 98 | 8.22 |
| DataONE\_CSDGM | CSDGM | 765 | 8.39 |
| GLEON | EML | 13 | 8.77 |
| USANPN | EML | 6 | 9.00 |
| CDL | CSDGM | 250 | 9.00 |
| PISCO | EML | 248 | 9.10 |
| ONEShare | EML | 109 | 9.50 |
| ESA | EML | 53 | 9.94 |
| DataONE | EML, CSDGM | 2818 | 10.35 |
| DataONE\_EML | EML | 2053 | 10.71 |
| TERN | EML | 250 | 11.00 |
| TFRI | EML | 250 | 11.17 |
| SEAD | CSDGM | 18 | 11.50 |
| LTER\_EUROPE | EML | 165 | 12.02 |
| KNB | EML | 167 | 12.50 |
| SANPARKS | EML | 247 | 12.59 |
| IOE | EML | 24 | 14.88 |
| KUBI | EML | 172 | 18.00 |

## Signature Sum Distribution

The signature sums averages show that CLOEBIRD and member nodes that use CSDGM have the most complete records on average. Do these member nodes have the best records? The following chart shows the signature sum for every record in the collection. CSDGM records are colored yellow, EML records are blue, and LTER records are orange. As we can see the most complete signature sums are all from LTER, and LTER records are mostly more complete than the average record, represented as a star.

## Signature Scores by Recommendation Level

Since the LTER recommendation has levels, the concepts in the preceding levels are implicitly contained in each successive level. Thus, any record with no missing Identification level concepts might be considered more complete than a record with the same signature sum, missing concepts in later levels. An example of why the stack model works for the LTER recommendation, would be that it is hard to access the correct record if identification and discovery levels are incomplete, even if all the records are 100% accessible.

To avoid conflating the completeness of records that are missing the same number of concepts from different levels or creating a weighting system to count concepts in the different recommendation levels differently the level’s signatures have been addressed separately. In each of the distribution visualizations, every record is represented. LTER records are orange, other EML records are blue, and CSDGM records are yellow. Signatures are figured from recommendation maximum rather than dialect maximum to avoid considering a CSDGM record recommendation complete when it is only dialect complete for the recommendation. These visualizations allow us to see how LTER records are distributed throughout the DataONE sample set.

### Identification Level

The majority of records are only missing three concepts. Only LTER has records that are complete. Most LTER records are missing two concepts. CSDGM records trend towards more complete than the mean.

### Discovery Level

Most records are missing two concepts. LTER records are the only records that are complete, though LTER also has records missing all of the concepts. No CSDGM record is missing more than 2 concepts.

### Evaluation Level

Most records are missing 2 concepts. LTER is joined by TFRI and GOA in contributing records that are complete.

### Access Level

The Access level is the first level to be complete in both dialects. The majority of the LTER records are complete. The majority of the CSDGM records are complete. Only records from CDL and SEAD are not.

### Integration Level

Again, only LTER contributed records to the sample set that are complete though they also have records that do not use concepts from the Integration level at all.

# Conclusions and Further Questions

### Observation 1

LTER uses every concept in the recommendation. No other DataOne member node’s collection sample contained every concept.

### Conclusion 1

LTER has the most complete collection coverage because it is the only collection to contain all concepts in the recommendation.

### Observation 2

86%of the LTER sample are in the top 17% most complete signature groups

### Conclusion 2

LTER record more likely to be more complete than a record from any other member node. If we look at the entire collection of EML, there are many signature score groups. In fact, many of the signatures are unique within the collection, while other signatures are over 10% of the entire DataONE sample set. If you pair the signature score record groups with the member nodes from DataONE, you can see that most the LTER records in the metadata-set are distributed towards complete signature scores. In the top 17% of the entire sample set, 86% of the LTER sampled records occur. The following visualization shows the collection convergence towards the top of the most complete metadata records in the DataONE sample. LTER is the only EML collection with records that are complete for the Identification level. LTER and CLOEBIRD are the only collections using EML that contain records that are complete with respect to the Discovery level. GOA, GLEON, KNB, SANPARKS, PISCO, TFRI, and LTER have records that are Evaluation level complete. Over half of the LTER collection is Evaluation level complete. GOA, GLEON, KNB, IOE, SANPARKS, TFRI, USANPN, and LTER all have records that are complete with respect to the Access Level. LTER is the only collection with a record that is Integration level complete. LTER is the only collection that has a shining example of each LTER Recommendation Level.

### Observation 3

LTER contributes most of the Shining Examples.

### Conclusion 3

LTER more familiar with concepts and how to document.

### Observation 4

By level, LTER does not have a higher completeness percentage than all other member nodes LTER is not more complete on unweighted average either

### Conclusion 4

LTER is not favored as highly as a collection that contains few moderately complete records. LTER is more complete than the average of all DataONE member nodes that use EML including itself.

### Observation 5

Homogeneity leads to more complete concepts in a collection. Collections that have a high degree of homogeneity are also more likely to contain more unused concepts

### Conclusion 5

Homogeneity can be bad for completeness. CDL and TERN are examples of this.

### Questions

It appears CSDGM collections are more complete with respect to LTER. This case is only made more strongly when the dialect limitations are handicapped to dialect maximums for the levels. What are the common concepts between LTER and the FGDC recommendation that likely informed the creation of the CSDGM collections?

What effect does time have on record completeness? The LTER sample set may all be from 2005. Would new records from succeeding years be more complete? By improving the sampler to return a sample set published in a specific year it is possible to study this.

Why is LTER the only collection complete with respect to the Identification Level, but the concept occurrence percentage average of ESA is higher? LTER is made up of many sites itself. Perhaps this makes the sample set heterogeneous in that some sites have more experience with creating complete metadata than others, as there are a significant portion of the LTER sample that are complete or missing one or two concepts. (86%) so the rest must not be very complete. Are there multiple occurrences of metadata completeness evolution through time in the member node LTER that can be documented by creating a sample set with collections made up of records published in a specific time period?

Even though LTER has the most complete records in the DataONE sampling, the collection as a whole is not the most complete for any concept? Or recommendation level? Why are there a large number of LTER signature sums in the sample set that are less complete than the average for DataONE and EML at DataONE? It seems obvious that some records produced at LTER benefit from the recommendation. Do not all of the metadata record producers get the same level of guidance and use the same infrastructure to create their metadata?

This is not the case. Metadata is created at a number of sites. Can these sites be treated as member nodes in a new analysis and show a stronger case for collection evolution towards completeness through community usage of a recommendation?

# 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).

Gordon, S., 2016. scgordon/CrosswalksWorkflow/wiki [WWW Document]. CrosswalksWorkflow GitHub Wiki. URL https://github.com/scgordon/CrosswalksWorkflow/wiki (accessed 3.16.17).

Mecum, B., 2015. Fix bug causing strange MN subdirectories to be made in results NCEAS/metadig@e2e89b6 [WWW Document]. NCEAS metadig GitHub - sample-metadata.py. URL https://github.com/NCEAS/metadig/commit/e2e89b63c90f00e72fa4f630deee6b8e4a8f7e2e (accessed 3.16.17).