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

Sean Gordon, Ted Habermann The HDF Group

Table of Contents

[Highlights 1](#_Toc475689778)

[Abstract 1](#_Toc475689779)

[Abbreviations 1](#_Toc475689780)

[Keywords 1](#_Toc475689781)

[Process 4](#_Toc475689782)

[Results 6](#_Toc475689783)

## **Highlights**

•

## **Abstract**

Considerable data analyses use automated workflows to ingest data from public repositories, and rely on data packages of high structural quality. The Long Term Ecological Research (LTER) Network now screens all packages entering its long-term archive to ensure completeness and quality, and to ascertain that metadata and data are structurally congruent, i.e., that the data typing and formats expressed in metadata agree with that found in data entities. The EML Congruence Checker (ECC) system is a component of the LTER Provenance Aware Synthesis Tracking Architecture (PASTA), and operates on data tables in packages described with Ecological Metadata Language using the EML Data Manager Library, written in Java. Checking is extensible for other data types and customizable via a template. Reports are retained as part of the submitted data package, and summaries here reflect the general usability of LTER data for a variety of purposes. On average in 2015, site-contributed data in the LTER catalog were 95% compliant (valid) with the current suite of checks.

## **Abbreviations**

* EML, Ecological Metadata Language;
* LTER, Long Term Ecological Network;
* ;

## **Keywords**

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

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

* Create Collection: Used the sampler.py script to obtain a sample of up to 250 metadata records for each memberNode.
* Python Script: <https://github.com/NCEAS/metadig/blob/master/sample-metadata.py>
* Raw Metadata: <https://github.com/NCEAS/metadig/tree/master/results>  
  Cleaning process: Normalize schema location and EML version in records, change directory names to dialect codes rather than dialect version
* Cleaned metadata collections:

<https://github.com/NCEAS/metadig/tree/master/collections/DataOne>

* The analysis was conducted using the Oxygen Developer GUI to operate a transform that reports the organization, record, xpath, and element content for each element that contains text in a collection of records. This is then run through a workbook called a QuickE.
* Transform used:   
  <https://github.com/tedhabermann/Transforms/commit/21261fc0db93f6ed62da91676433876162dc1f3d#diff-f5eb263f14bbfaae7abc4faa89004508>
* Example of resultant data that plugs in to a QuickE:

<https://github.com/NCEAS/metadig/blob/master/contentDistributionFiles/EML/EML2.1.1/metadataContent.EML2.1.1.txt>

* Data in QuickE Workbook:

<https://github.com/NCEAS/metadig/blob/master/contentDistributionFiles/EML/EML2.1.1/Content.QuickE.EML.2.1.1.xlsx>

* I then iterated over the other dialects with larger holdings in DataONE and put the content of the text file into the data worksheet in the Content.QuickE template.
* Template: <https://github.com/NCEAS/metadig/blob/master/contentDistributionFiles/Template/Content.QuickE.xlsx>
* Resultant workbooks and the original data sheets:   
  <https://github.com/NCEAS/metadig/tree/master/contentDistributionFiles/CSDGM>  
  <https://github.com/NCEAS/metadig/tree/master/contentDistributionFiles/Dryad>  
  <https://github.com/NCEAS/metadig/tree/master/contentDistributionFiles/DublinCore>  
  <https://github.com/NCEAS/metadig/tree/master/contentDistributionFiles/EML>
* Further work was done in Excel to refine the analysis to focus on elements in the FGDC recommendation and the first three levels of the LTER recommendation.

FGDC:  
<https://github.com/NCEAS/metadig/blob/master/contentDistributionFiles/FGDCrec_FGDC%2BEMLdialects.xlsx>  
LTER:   
<https://github.com/NCEAS/metadig/blob/master/contentDistributionFiles/LTER3lvl_FGDC%2BEMLdialects.xlsx>

## Results

