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

Sean Gordon, The HDF Group

Abstract 1

Introduction 1

Metadata Standards/Dialects/Recommendations 1

Recommendations and Dialects at DataOne 2

LTER Recommendations 3

Data 3

DataOne Member Nodes 3

Sampling Metadata Collections 3

Results 3

Conclusions 3

Process 5

# Abstract

Abstract goes here

# Introduction

All scientists and scientific communities recognize the need to document observations and processing clearly and completely in order 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 in notebooks, scientific papers, web pages, user guides, word processing documents, spreadsheets, data dictionaries, PDF’s, databases, custom binary and ASCII formats, and almost any other 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 particular 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, helps address discovery, access, use, and understanding by providing 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 particular 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

Most scientific communities recognize the need for metadata and address that need using one of several approaches: they 1) either use a metadata standard proposed by a related community or organization, or 2) they develop one that fits their perceived needs. These metadata standards include standard concept names, definitions and associated structures. 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*.

Recommendations and dialects are different but, in practice, they are frequently related because groups create metadata dialects and recommendations together. For example, the NASA Global Change Master Directory (GCMD) created a metadata dialect for exchanging directory information (Directory Interchange Format, DIF) schematically shown as D1 in Figure 1. When GCMD created DIF, they also created recommendations at three levels (mandatory, recommended, and suggested). These recommendations (R1, R2, and R3) contained different numbers of elements (mandatory recommendations are typically the smallest). All of these elements were entirely contained in the DIF dialect because they were created as guidance for how to use DIF. This is a typical situation: a dialect with multiple dialect-specific recommendations.

At some other time, another organization created a different dialect (D2) with two other recommendations (R4 and R5), also completely contained by their dialect (D2). R4 also had considerable overlap with D1 and also with R1. This is not unusual when two organizations are addressing the same metadata use case (e.g. Discovery).

Finally, another organization created one more recommendation (R6) that was independent of dialect (i.e. purely conceptual). This recommendation included concepts that were also included in D2 and R5. A real-world example of this is the Open Geospatial Consortium (OGC) Catalog Service for the Web (CSW) recommendation(s). These were developed specifically to be dialect-independent with profiles defining their implementations in specific dialects, e.g. CSDGM and ISO 19115.

## 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 Ecological Metadata Language (EML) and the Content Standard for Digital Geographic Metadata (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

### Sampling Metadata Collections

# Results

# Conclusions

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. What is quality metadata though? How can you be sure that what you include in the metadata is not only relevant to your organization’s collections, but understandable to your scientific community and beyond?

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 particular form of the documentation language that is specific to a community.

|  |
| --- |
| Known Metadata Dialects in DataONE |
| ISO 19115-1 / ISO 19115-3 |
| Directory Interchange Format (DIF9) |
| Directory Interchange Format(DIF10) |
| DataCite 3.1 |
| Content Standard for Digital Geospatial Metadata (CSDGM) |
| Content Standard for Digital Geospatial Metadata (CSDGM) Biological Data Profile |
| ECHO |
| ISO 19115 and ISO 19115-2 / ISO 19139 and ISO 19139-2 |
| Metadata Object Description Schema (MODS) |
| Dryad |
| Dublin Core Extended v1.0 |
| 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 |
| UMM-Collection | National Aeronautics and Space Administration |
| UMM-Common | National Aeronautics and Space Administration |
| UMM-Granule | National Aeronautics and Space Administration |
| LTER\_Completeness | Long Term Ecological Research Network |
| Dryad-Package | Dryad Digital Repository |
| Dryad-File | Dryad Digital Repository |
| ISO-1\_Access | International Standards Organization |
| SERF\_Access | National Aeronautics and Space Administration |
| ISO\_Lineage\_Understanding | International Standards Organization |
| ISO\_Quality\_Understanding | International Standards Organization |
| WSDL\_Access | World Wide Web Consortium |
| ISO\_Reference\_Understanding | International Standards Organization |
| ISO\_UserFeedback\_Understanding | International Standards Organization |

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 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, comparing across the Member Nodes using CSDGM or EML.

## Process

Additional presentations:  
<https://github.com/NCEAS/metadig/blob/master/contentDistributionFiles/Presentations/Powers_DataONE_AHM_2015.pptx>

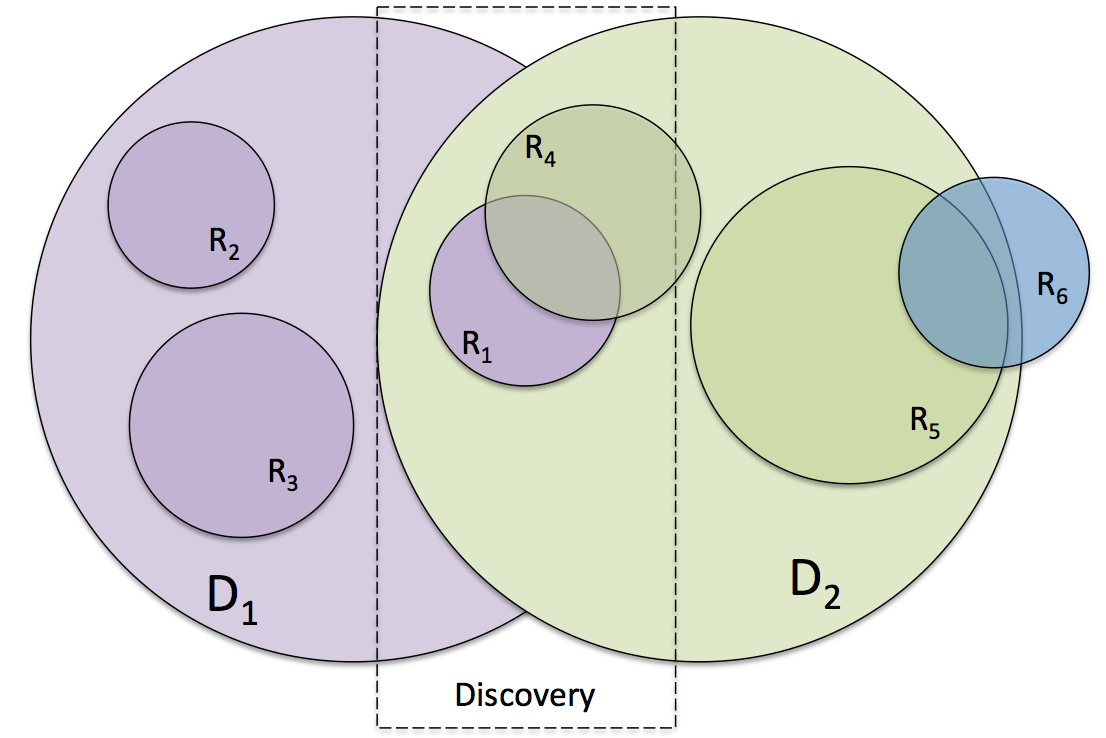


Figure 1. Metadata Dialects and Recommendations