# Is Your Metadata Catalog in Shape? A Critical Assessment of Metadata Quality, Structure, and Integrity in Digital Health and Scientific Information Systems

Dr. Lukas Weber[1]*
[1] University of Copenhagen, Department of Health Informatics and Biomedical Data Management, Copenhagen, Denmark

## Introduction

It can be difficult to maintain the fitness of a metadata catalog. Catalogs are increasingly diverse and rapidly growing in size. Without good metadata it is difficult for users to discover datasets likely to be useful, but more importantly good metadata is essential for converting that dataset into information so it can be transformed into knowledge and understood and used in new and novel ways.

How can curators ensure the catalog's ability to meet the different needs of the users of data? How can catalogers quickly evaluate the likelihood that a record in the catalog will contain all of the metadata needed by a potential user of any resource described, and which collections need specific types of improvement? A visualization that can provide these insights is needed that can impart the shape of the catalog's fitness for meeting different user's information needs.

## A metadata metaphor

We can think of metadata as a key to unlock information from data's door. Keys have a shape to fit a specific lock. We can think of this lock as a community's information needs. A good example of expressed information needs would be a metadata standard's recommendation or a community's stated best practices, such as the EML Best Practices for LTER Sites (LTER, 2004). Each pin in a locks cylinder can be a specific need such as a title or abstract, even something as complex as methods or as specific as what a column in a table means.

Different communities and organizations will have different needs to unlock information and the visualization will work for any of them, be it discovery or understanding, or to ensure that credit for unique contributions goes where it is due, as with software and data citation efforts.

The lock's cylinder pins can be as simple as a test for presence/absence of an element in the structure of a document such as the tests used to compare EML and CSDGM producing member nodes in DataONE for Best Practices completeness to determine if community recommendations can improve metadata completeness (Gordon; Habermann, 2018).

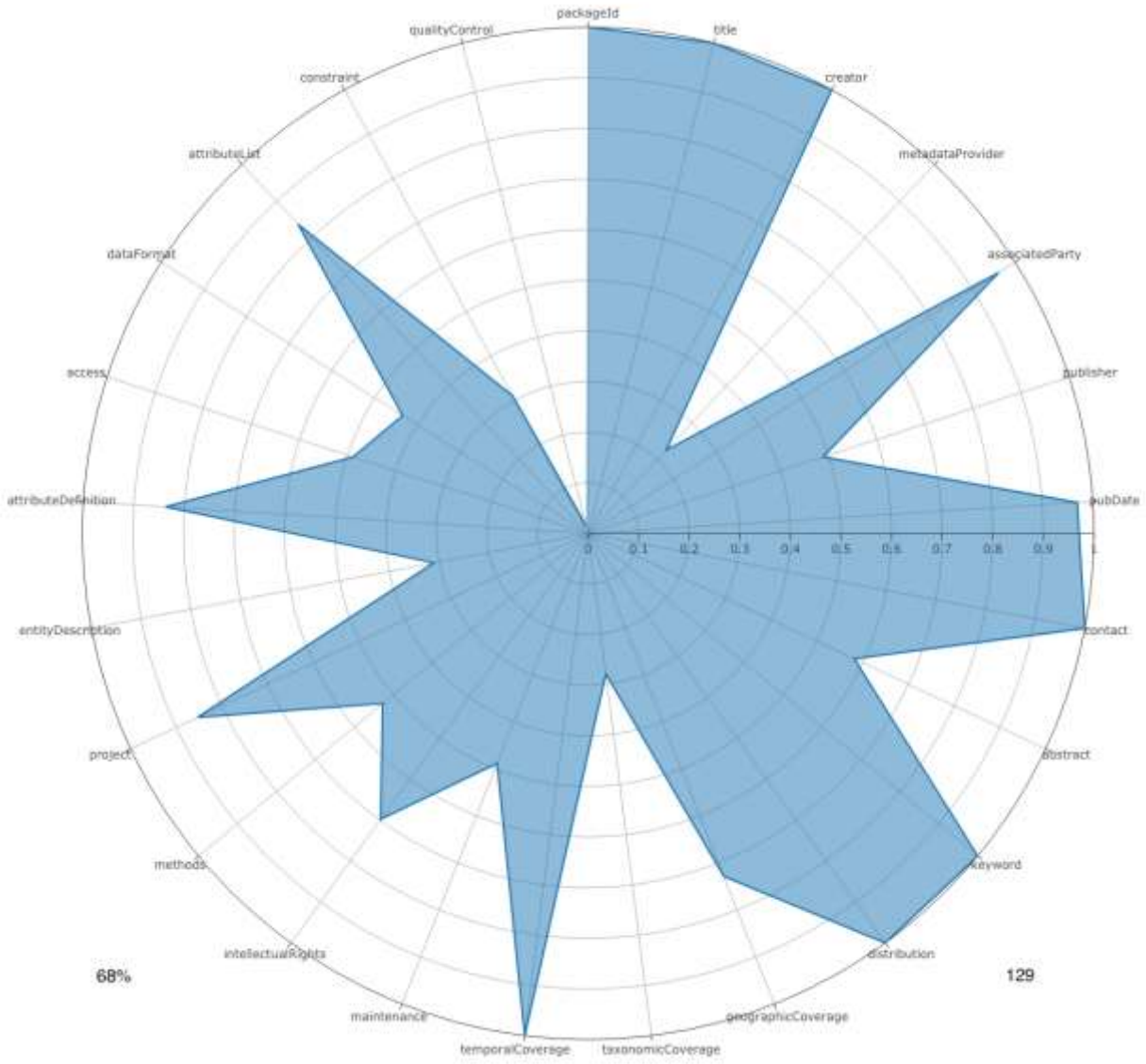
Tests can also be a more complex quantifiable that evaluates the content of that element such as with LTER's Pasta (Cite, 2011) or NCEAS's Metadata Quality Engine (Cite, 2019). Any test can

provide a completeness percentage for the collection by aggregating the results for each document.

## The shape of fitness

To visualize a metadata catalog or collection's ability to meet the information needs of a community, plot each element as an axis in a polar chart and then quantify from 0 to 1 how many records pass whatever test desired to evaluate the element (Gordon, 2019). Figure 1 is an example of the visualization of an information need using the EML Best Practices for LTER Sites elements. The shape expressed in Figure 1 is calculated using a collection of records from an LTER site.

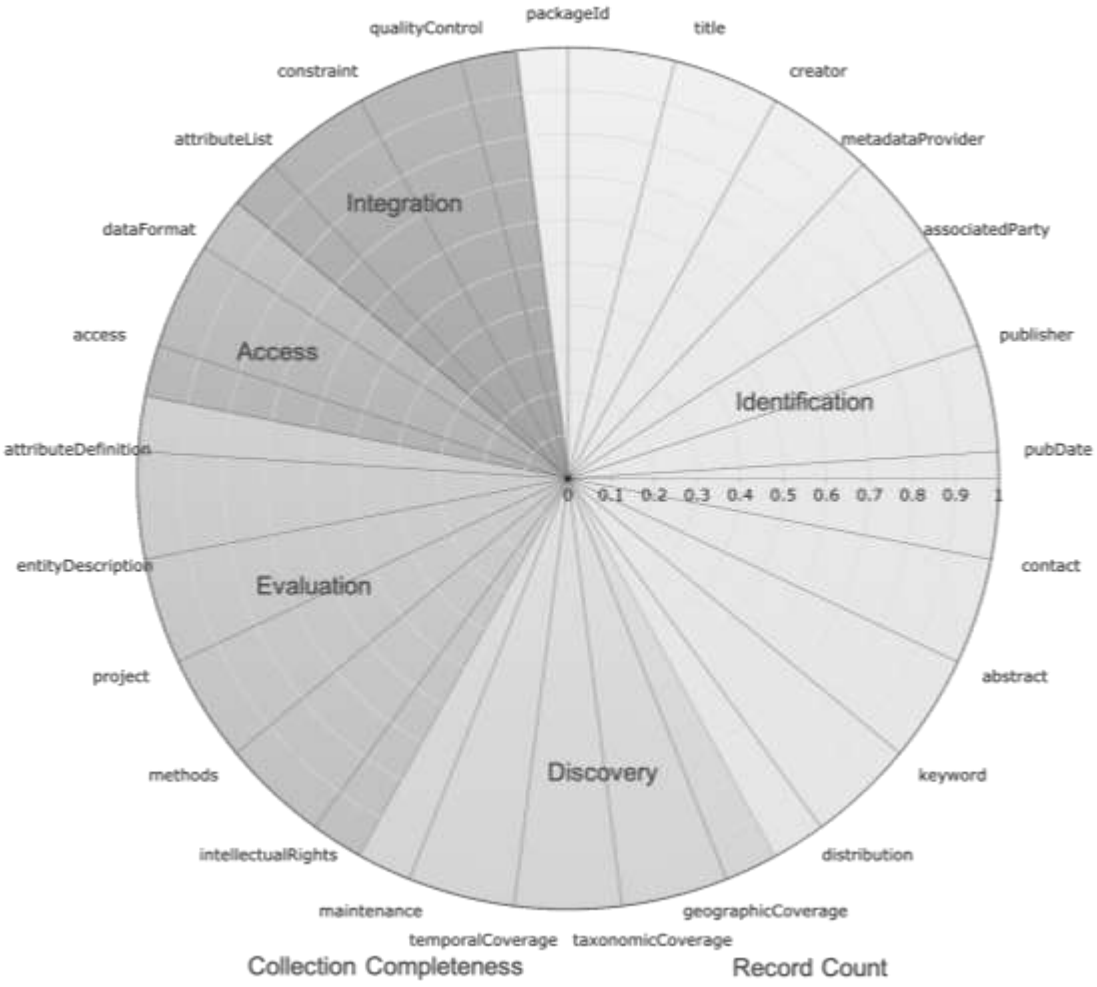
FIGURE 1



## Many use cases, one community

Often, as is the case with LTER's Best Practices, a community's information need can be broken up into different needs or use cases. The EML Best Practices for LTER Sites has 5 different use cases they call levels. They are Identification, Discovery, Evaluation, Access, and Integration. These levels can be plotted within a single chart as with Figure 2 (Gordon; Habermann, 2019) or separately depending on focus of inquiry.

FIGURE 2




Applying this visualization style to a collection as it changes through time, like in Figure 3 can provide additional insight such as changes in information needs over time, and chronicle the improvements and successes in providing users with a complete catalog (Gordon, 2019).

FIGURE 3

## Different queries, different insights

Some elements are only used when the information is present. The FGDC community's remarkable recommendation and guidance documentation for the CSDGM (FGDC, 1998) contains many such structural requirements for more specialized metadata concepts and consider them mandatory if applicable to the resource being documented. Other communities are less explicit about varying priorities for elements.

While this lowers the overall percentage of a complete information need for the collection, it is not necessarily indicative of a shortcoming, it is simply a more unique shape of completeness than the basic full circle, and may indicate a need for refining the recommendation. The effect of different prioritization of use cases can be minimized by providing a shape for each level or discrete part of the recommendation by focusing on metadata for specific use cases in separate charts.

By applying the visualization to every element used in the catalog or collection, insights as to which metadata concepts are actually important to the specific members of a community and a data-driven best practices for essential metadata can be derived.

In many metadata standards in earth science there are multiple elements to give structure to the same metadata concepts. Different record providers will use slightly different elements to express the same kinds of information even within the same community. This can be observed and potentially used to normalize element usage which will aid the transition from legacy documentation formats and standards to linked data vocabularies with more transparent definitions.

## Transforming data into knowledge

This visualization is a way that repositories and other maintainers of metadata catalogs can proactively stay on top of the completeness and quality of the records they offer, without needing to check each record over and over as new information needs develop, or worse, wait until problems are discovered by users. It will facilitate their efforts to aid users in their quest for relevant data and ensure they are able to transform the data into information, furthering their own edge of knowledge.

## Bibliography


- EML Best Practices Document 2004. (2004, October 29). Retrieved February 7, 2019, from https://lternet.edu/documents/eml-best-practices-2004/
- Gordon, S., & Habermann, T. (2018). The influence of community recommendations on metadata completeness. *Ecological Informatics*, *43*, 38–51. https://doi.org/10.1016/j.ecoinf.2017.09.005
- Slaughter, P., Leinfelder, B., Mecum, B., Gordon, S., Jones, M., & Mullen, D. (2019). *MetaDig Engine*. Java, National Center for Ecological Analysis and Synthesis. Retrieved from https://github.com/NCEAS/metadig-engine (Original work published 2016)
- Gordon, Sean. (2019, January 1). LTER Sites: Evaluate, Analyze, Report (Version v0.0.4). Zenodo. http://doi.org/10.5281/zenodo.2539415
- Gordon, Sean, & Habermann, Ted. (2019). Visualizing The Evolution of Metadata (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.2538983
- Geospatial Metadata Standards — Federal Geographic Data Committee. (1998). Retrieved June 23, 2014, from https://www.fgdc.gov/metadata/geospatial-metadata-standards