Improving Metadata with Automated Quality Evaluation

Bryce Mecum
Scientific Software
National Center for Ecological Analysis and Synthesis
@brycem
orcid.org/0000-0002-0381-3766
mecum@nceas.ucsb.edu
The Team

Matthew B. Jones
Peter Slaughter
Ben Leinfelder
Bryce Mecum

Ted Habermann
Lindsay Powers
Sean Gordon
Overview

- Metadata is great, when present
- Some metadata records are better than others
- It really depends on your purpose

- There is an interaction between the metadata and the community using it
- Many communities have already established what qualifies as good metadata
Metadata exists on a continuum

We often find ourselves around here

No metadata  Minimal metadata  Pretty good metadata

And we're trying to move over here
Existing Recommendation

Attribute Convention for Data Discovery (ACDD)
http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery

Highly recommended: Title, summary, keywords, attribute name/units, etc...
Recommended: identifier, creator name+email, basic spatial/temporal bounds

much more
Existing Recommendation

LTER PASTA Quality Suite

Example Checks:

• Data can be loaded into a relational database
• CSV field delimiter matches data
• # header lines matches data
• ... 32 in all
Overview

Better metadata is important
• We build our search portals around it
• We need it to re-use data
• We need it to understand data

Metadata can improve at multiple stages
• When the metadata are being authored
• At metadata/data submission time
• After-the-fact (collection level)
Metadata Quality Engine

• Automatically grade metadata records
• Supports the types of checking communities already do
• Can be deployed alongside existing software

• Target audiences:
  • Producers (Individual Researchers)
    • At metadata/data submission
  • Data repositories
    • At the collection level
  • Consumers (Individual Researchers)
    • At record level, for use and interpretation
Metadata Quality Engine

- Supports any XML-based metadata standard
  - ISO19115
  - EML
  - FGDC
  - etc...

- Write Checks in the same language you do your science in

Languages:
- R
- Python
- Java
Architecture

Metadata

Data

Metadata Quality Engine

Recommendation

Title Check

Units Check

Methods Check

...

Quality Report
## Recommendations

- Collection of *Checks*, like unit tests for metadata/data
- Community-oriented
  - Can mix and match *Checks* in other *Recommendations*
  - Or write your own

<table>
<thead>
<tr>
<th>Check Name</th>
<th>Check</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive Title</td>
<td>Title exists, &gt; 7 words</td>
<td>Metadata</td>
</tr>
<tr>
<td>Unique Attribute Names</td>
<td>Attribute names are unique within each entity</td>
<td>Metadata</td>
</tr>
<tr>
<td>Valid Units</td>
<td>Units are all from a controlled vocabulary</td>
<td>Metadata</td>
</tr>
<tr>
<td>Schema Valid</td>
<td>Metadata validates according to its schema</td>
<td>Metadata</td>
</tr>
<tr>
<td>Checksum Matches</td>
<td>Data checksums match metadata</td>
<td>Congruency</td>
</tr>
<tr>
<td>Data Links Live</td>
<td>All URLs return content</td>
<td>Congruency</td>
</tr>
<tr>
<td>Duplicate Data Rows</td>
<td>Get a count of duplicate data rows</td>
<td>Data</td>
</tr>
</tbody>
</table>
Matthew Shupe. 2016. Sodar measurements. NSF Arctic Data Center. doi:10.18739/A2XW9X.

Files in this dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>File type</th>
<th>Size</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata: Sodar measurements</td>
<td>EML v2.1.1</td>
<td>7 KB</td>
<td>4 views</td>
</tr>
</tbody>
</table>

General

Identifier: doi:10.18739/A2XW9X

Abstract

This data set contains raw measurements from a sodar that is deployed at Summit Station, Greenland. This sodar operates at 2100 Hz and points vertically. Reflectivity measurements from the system are used to characterize gradients in the lower atmosphere that provide information on the boundary layer depth. Data files including the relative backscatter. Detailed information on file parameters and other aspects of the dataset are included in the netCDF header information for each file. The sodar is owned by the National Oceanic and Atmospheric Administrations's Earth System Research Laboratory. Field operations are supported by the National Science Foundation's Arctic Observing Network (AON) Program.

Publication Date: 2016-10-06
Metadata Quality Report

Matthew Shupe. 2016. Sodar measurements. NSF Arctic Data Center. doi:10.18739/A2XW9X.

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.

- **Identification**: 91% complete
- **Discovery**: 100% complete
- **Interpretation**: 100% complete

- Passed 16 checks out of 18. Good job!
- Warning for 1 check. Please review these warnings.
- Failed 1 check. Please correct these issues.
- 8 informational checks. These may include skips, errors and failures.
### Passed 16 checks out of 18. Good job!

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Status</th>
<th>Identification</th>
<th>Requirement</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>At least one award number was found.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>All award numbers were found in the NSF award database.</td>
<td>✔️</td>
<td>✔️</td>
<td>OPTIONAL</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>One creator is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>One contact is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>The abstract is 114 word(s) long which is sufficient.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>An identifier is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>The identifier looks like a DOI.</td>
<td>✔️</td>
<td>✔️</td>
<td>OPTIONAL</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>The document is licensed with a Creative Commons CC-BY license.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>A description of this dataset's temporal coverage is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>A textual description of the geographic coverage of this dataset is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>A set of bounding coordinates describing the geographic coverage of this dataset is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
<tr>
<td>A publication date is present.</td>
<td>✔️</td>
<td>✔️</td>
<td>REQUIRED</td>
<td>SUCCESS</td>
</tr>
</tbody>
</table>
## Warning for 1 check. Please review these warnings.

- No data descriptions are present.

## Failed 1 check. Please correct these issues.

- The number of words in the dataset's title is 2. The minimum required word count is 7.

## 8 informational checks. These may include skips, errors and failures.

- All creators have email addresses. All creators have addresses.
- All contacts have email addresses. All contacts have addresses.
- No data descriptions are present, so unable to check entity 'name', 'format', etc.
- No data table descriptions (and related attributes) are present.
- No data table descriptions are present, so cannot check attribute definition word counts.
- No data table descriptions are present, so cannot check if attribute names and definition differ.
Metadata Quality Engine

• Operates across metadata standards
• Can check metadata, data, and the references between the two
• Uses a REST API to separate the Engine from what's being checked
• Quality Reports can be indexed to compare records
Products to date

- Project materials: https://github.com/NCEAS/metadig
- Quality Engine https://github.com/NCEAS/mdqengine
- Web app https://github.com/NCEAS/mdq-webapp
- Three Recommendations
  - Arctic Data Center (https://arcticdata.io)
  - LTER PASTA
  - CSW Core Queryables
- Integrated into our existing repository software (Metacat)
- Deployed on Arctic Data Center (https://arcticdata.io)
Challenges

• What to show the user and how
  • Percent / ratio (e.g., 17/20 Checks passed)
  • Percentile (how do I compare to others?)
Thanks!

Bryce Mecum  
@brycem  
orcid.org/0000-0002-0381-3766

Ted Habermann  
@TedHabermann  
orcid.org/0000-0003-3585-6733

Matthew B. Jones  
@metamattj  
orcid.org/0000-0003-0077-4738

This work was funded by NSF award 1443062.