

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



NCEAS

National Center for Ecological Analysis and Synthesis

DataONE

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

Citation

Title present
Pub. date present
Author(s) present

Getting the data

Landing page present
Service description(s) present

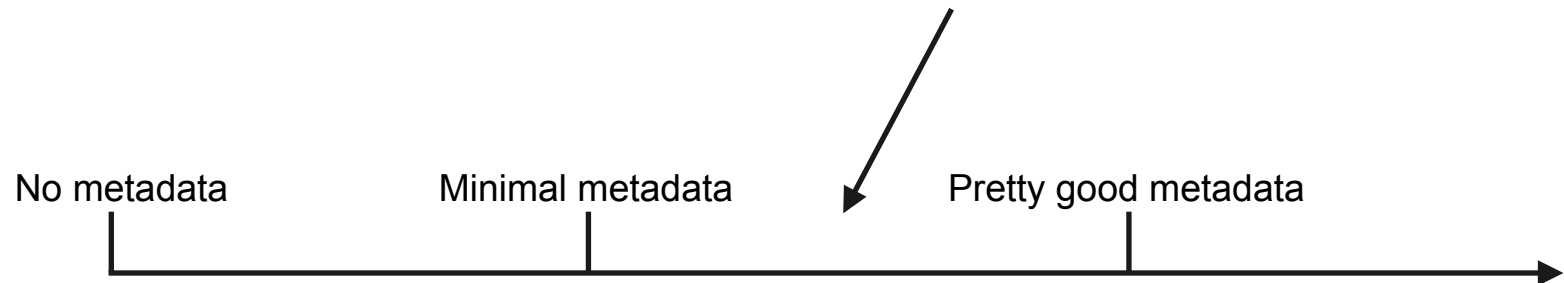
Using the data

Methods present
Variables defined w/ units/scales, etc

- 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



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

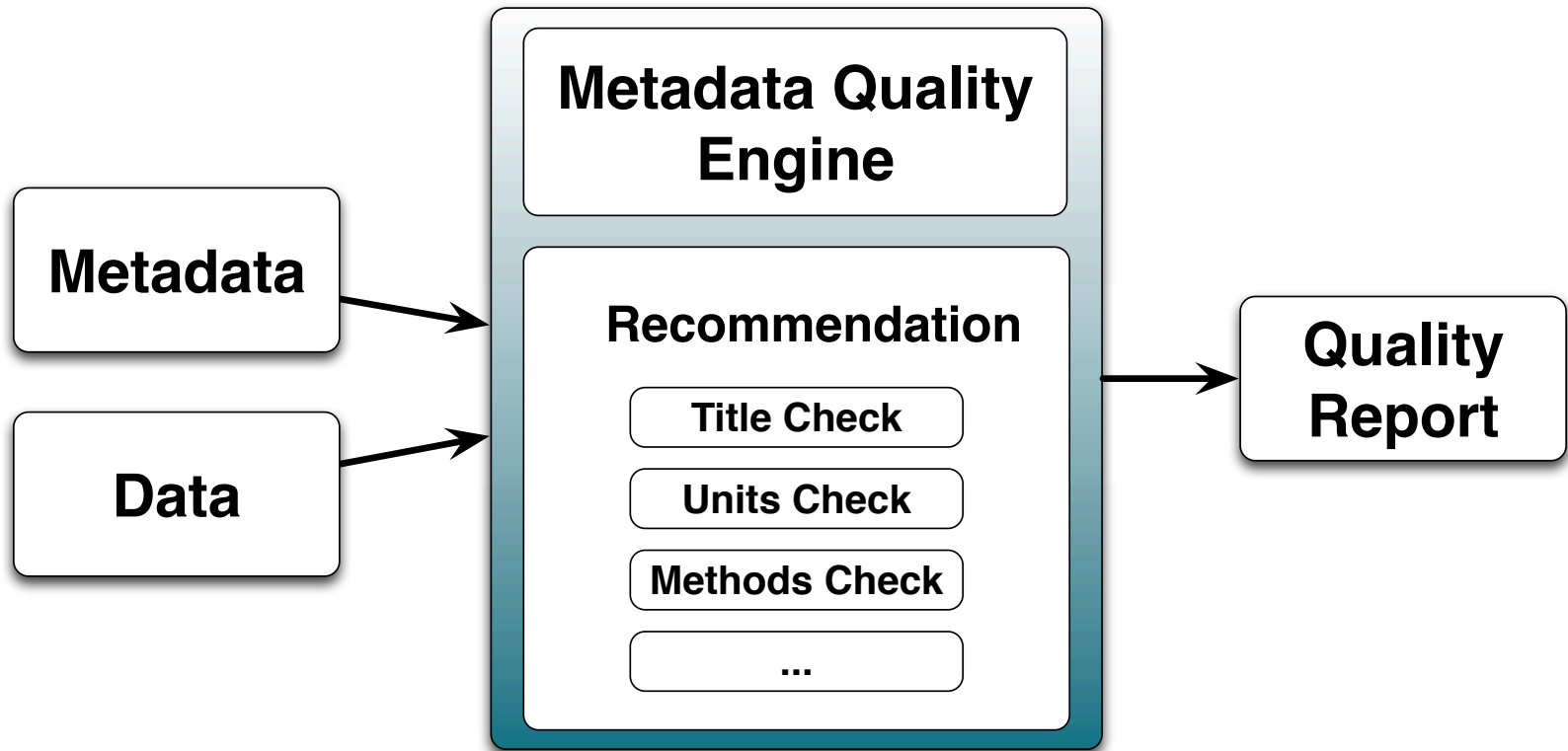
- Supports any XML-based metadata standard



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



Architecture



Recommendations

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





Check Name	Check	Type
Descriptive Title	Title exists, > 7 words	Metadata
Unique Attribute Names	Attribute names are unique within each entity	Metadata
Valid Units	Units are all from a controlled vocabulary	Metadata
Schema Valid	Metadata validates according to its schema	Metadata
Checksum Matches	Data checksums match metadata	Congruency
Data Links Live	All URLs return content	Congruency
Duplicate Data Rows	Get a count of duplicate data rows	Data

[Home](#) / [Search](#) / [Metadata](#)

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

[Copy Citation](#)
[Quality report](#)


Files in this dataset Package: resource_map_doi:10.18739/A2XW9X

	Name	File type	Size	Downloads	Download all 
	Metadata: Sodar measurements	EML v2.1.1	7 KB	4 views	Download 

General

Identifier

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.

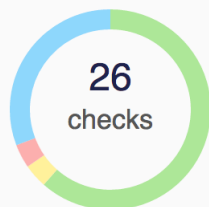
Publication Date

People and Associated Parties

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!

✓	At least one award number was found.	?	identification	REQUIRED	SUCCESS
✓	All award numbers were found in the NSF award database.	?	identification	OPTIONAL	SUCCESS
✓	One creator is present.	?	identification	REQUIRED	SUCCESS
✓	One contact is present.	?	identification	REQUIRED	SUCCESS
✓	The abstract is 114 word(s) long which is sufficient.	?	discovery	REQUIRED	SUCCESS
✓	An identifier is present.	?	identification	REQUIRED	SUCCESS
✓	The identifier looks like a DOI.	?	identification	OPTIONAL	SUCCESS
✓	The document is licensed with a Creative Commons CC-BY license.	?	identification	REQUIRED	SUCCESS
✓	A description of this dataset's temporal coverage is present.	?	discovery	REQUIRED	SUCCESS
✓	A textual description of the geographic coverage of this dataset is present.	?	discovery	REQUIRED	SUCCESS
✓	A set of bounding coordinates describing the geographic coverage of this dataset is present.	?	discovery	REQUIRED	SUCCESS
✓	A publication date is present.	?	identification	REQUIRED	SUCCESS

▼ Warning for 1 check. Please review these warnings.

 No data descriptions are present.  interpretation OPTIONAL FAILURE



▼ 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.  identification REQUIRED FAILURE


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



 All creators have email addresses.
All creators have addresses.  identification INFO SUCCESS

 All contacts have email addresses.
All contacts have addresses.  identification INFO SUCCESS

 No data descriptions are present, so unable to check entity 'name', 'format', etc  interpretation INFO SKIP

 No data table descriptions (and related attributes) are present.  interpretation OPTIONAL SKIP

 No data table descriptions are present, so cannot check attribute definition word counts.  interpretation OPTIONAL SKIP

 No data table descriptions are present, so cannot check if attribute names and definition differ  interpretation OPTIONAL SKIP

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.

