CMR Metadata Curation

Winter ESIP 2017
Dana Shum
Kaylin Bugbee

The material is based upon work supported by the National Aeronautics and Space Administration under Contract Number NNG15HZ39C
We have metadata problems

- Providers names have been genericized to protect the innocent
- Analysis run on BEDI collection metadata in ESDIS' Common Metadata Repository (CMR) on 11/11/16.
- Unified Metadata Model for Collections (UMM-C) schema can be found here: https://git.earthdata.nasa.gov/projects/EMFD/repos/unified-metadata-model/browse
The problems are visible to users

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM-1</td>
<td>48</td>
</tr>
<tr>
<td>TERRA</td>
<td>1</td>
</tr>
<tr>
<td>SPECTRAL ENGINEERING</td>
<td>2</td>
</tr>
<tr>
<td>SPECTRAL/ENGINEERING</td>
<td>548</td>
</tr>
<tr>
<td>BIOOSPHERE</td>
<td>2</td>
</tr>
<tr>
<td>BIOSPHERE</td>
<td>830</td>
</tr>
</tbody>
</table>

- Incorrect scientific information
- Missing required fields
- Invalid Keywords
- Outdated contact information
- Inconsistent use of terms
- Misspellings

*The list goes on...*
The Plan to Fix It!

- **Kickoff**
  - Brief Mgmt
  - Identify POCs

- **ESDIS DAAC Collaboration**
  - Metadata Summits
  - Report Generation

- **Collaborative Curation**
  - Correct identified issues
  - Resubmit corrected metadata

- **Curation Maintenance**
  - Continually evolve metadata to meet changing guidance

- **The Team**
  - Analysis and Review of CMR (ARC) Team
    - Based at Marshall Space Flight Center
    - Team comprised of Earth Science data specialists
    - Team focuses on CMR metadata review
  - CMR Team
    - Team comprised of Software and Systems Engineers
    - Team focuses on UMM-C model enhancements and programmatic metadata validation
  - CMR and ARC teams have collaborated with the GCMD to develop and document unified metadata curation guidelines and best practices
What will be looked at?

• The ARC Team will contact the assigned POC at the beginning of each ESDIS DAAC’s review process
• All collection level records will undergo review:
  – Each record will be reviewed for correctness, completeness and consistency
  – Both an automated and manual reviews will be conducted for each record
  – For each collection record, a randomly selected granule level record will also be reviewed
  – Native metadata format will be reviewed

Rules for metadata checks can be reviewed here (Note: Rules document is a living document):
https://wiki.earthdata.nasa.gov/pages/viewpage.action?pageId=77398170
Tools to Help

- Metadata Mgmt Tool (MMT)
- CMR API
- DocBuilder
- Documentation
https://mmt.earthdata.nasa.gov

METADATA MANAGEMENT TOOL (MMT)
MMT Visual Validation

The icons below each form name indicate progress toward completion. See below for information on what each icon represents.

- Required but not complete
- Required and complete
- Optional and not complete
- Optional and complete
- Does not pass validation

Metadata Fields

Collection Information
Descriptive Keywords
Spatial Information
Collection Citations
Organizations
Metadata Information

EOSDIS

8

SESIP-0716-TBD
Preview Panel / Collection HTML
Keyword Validation

![Keyword Validation Image](image-url)
Help Text

Platform-specific characteristics, e.g., Equator Crossing Time, Inclination Angle, Orbital Period. The characteristic names must be unique on this platform; however, the names do not have to be unique across platforms.

Validation
- Minimum Items: 0
https://cmr.earthdata.nasa.gov

CMR API SUPPORT
CMR Validation

The CMR exposes a Validation endpoint so that providers can test their metadata prior to ingesting it. The same validations are applied there as on the ingest endpoint.

**Native XML Validation** - Collections are validated against their native schemas (ECHO10, DIF9, DIF10, ISO-19115-1, ISO-19115-2)

**UMM-C Validation** - Collections are converted from their native formats into the Unified Metadata Model for Collections (UMM-C). Field formats, ranges and controlled vocabularies are validated. This optionally includes validation of GCMD Keywords.

**Business Rule Validation** - Collections are validated against their existing granules to ensure integrity is maintained.
CMR Business Rule Validation

• Delete time in collection metadata is after the current time.
• Collection additional attribute changes do not invalidate existing granules.
• Collection projects(campaigns) changes do not invalidate existing granules.
• Collection temporal changes do not invalidate existing granules.
• Collection spatial changes do not invalidate existing granules.
Humanizers (a.k.a. Bandaids)

- Humanizers are a set of instructions, or aliases, which allow CMR administrators to quickly clean-up the faceted display of metadata. They **do not** change the underlying metadata.
  - Misspellings: “Bioosphere”
  - Legacy Terms: “AM-1” instead of Terra
  - Inconsistent Names: Processing levels “Level 1”, “1”
  - Remove whitespace around words
  - Keyword Case Corrections: Use normal case when appropriate
Tags (Enhanced Metadata)

• Tags allow layering on additional information to one or more collections.

• Potential Curation Uses
  – Key
    • “provider.extra.curationStatus”
  – Description
    • “Used to store the curation status of provider’s curation efforts”
  – Category (optional)
    • “Curation”

• Tags can be retrieved and searched
Useful Documentation

- **CMR Curation Home:**
  [https://wiki.earthdata.nasa.gov/display/CMRARC/CMR+Metadata+Curation+Home](https://wiki.earthdata.nasa.gov/display/CMRARC/CMR+Metadata+Curation+Home)

- **ARC’s Rules for Metadata Checks:**

- **UMM-C documentation:**
  [https://wiki.earthdata.nasa.gov/display/CMR/CMR+Documents](https://wiki.earthdata.nasa.gov/display/CMR/CMR+Documents)

- **GCMD DIF Metadata Writer’s Guide:**

- **#Curation Slack Channel:**
  [https://eosdis.slack.com/messages/curation/](https://eosdis.slack.com/messages/curation/)

- **ESIP Wiki**
A CASE STUDY ON TRYING TO DO THE RIGHT THING
We heard that we should use Terra instead of AM-1. We want to fix it! Let's try using MMT, we've heard it's great!

Can't wait to help you! While you're here though, I'm going to need you to also fix the other 30 errors with each of your records before I can submit them.

Whoah. A bit more than we had planned on. Plus, we need GCMD Keywords that don't even exist yet!

We worked with GCMD and got those keywords added. But if I ingest your collection now, you'll invalidate your millions of granules. Let us know when you are ready to fix all of your granules.

OK, that's a lot of work. Any chance you can modify those validation rules to help us out in the near term?

Yup, I think we can apply humanizers to that problem. Working on it now...
Some Early Lessons

• This is a collaborative effort between a lot of different teams.
• Communication is critical.
• Tools will need to be tweaked. Often.
WHAT ARE WE MISSING?
Help us!

• **What other tools would be useful?**
  – Bulk updates?
  – Rubrics?
  – ?

• **What documentation would be useful?**
  – Fully filled out *best* examples of each format?
  – ?

• **Preferred Communication mechanisms?**

• **Metadata Specifics**
  – How/where to specify DOIs?
  – Should Citation information be stored in pieces or in one combined free form string?
Questions?

• Metadata Review / Rules Questions:
  – Contact Kaylin Bugbee - kaylin.m.bugbee@nasa.gov

• Tool Related Questions:
  – Contact Dana Shum – dshum@raytheon.com
This material is based upon work supported by the National Aeronautics and Space Administration under Contract Number NNG15HZ39C.

Raytheon