Analysis and Review of NASA Earth Science Metadata

How Automation Plays a Role

Patrick Staton[1], Jeanne le Roux[1], Kaylin Bugbee[1], Alicia Aleman[2], and Thomas Cherry[2]

1 – University of Alabama in Huntsville; 2 – EED-2/SGT Inc.

IMPACT
Interagency Implementation and Advanced Concepts Team

What is ARC?

The Analysis and Review of the Common Metadata Repository (CMR ARC) Team reviews all EOSDIS metadata. The team's objective is to achieve consistency, correctness, and completeness for all metadata records in the CMR, as well as improve the discoverability of NASA's Earth Science data within the CMR framework. This work is currently being completed at Marshall Space Flight Center.

CMR makes a single discovery point possible for NASA's Earth Science data users. The CMR team, in collaboration with three other core metadata teams, contributes to the stewardship of NASA's Earth Science data through a process of continual curation and the ongoing development of the Unified Metadata Model (UMM).

A key tool now used in the curation process, referred to as the NASA CMR Dashboard, is an online curation dashboard developed in collaboration with software development company, Element 84. This tool facilitates the review of Earth Science metadata records and subsequent stakeholder collaboration on the resolution of identified issues. A key capability of the new tool is a suite of automated compliance checks written in Python 3.6 that verify the integrity of various metadata elements across multiple standards.

Methods

The ARC team's method of metadata evaluation consists of three processes:

I. Evaluate and assess metadata

II. Provide recommendations to improve overall quality of the metadata

III. Submit recommendations to the appropriate DAAC for implementation

For some elements, the CMR ARC team is only concerned with the presence of a value; whereas other elements need to be scrupulously validated against an EOSDIS standard-specific schema. The automated compliance checks include the testing of logical collection-granule relationships, the handling of URL HTTPS response codes, the validation of controlled keywords, and more.

python™

Automated Metadata Analysis

There are three different metadata standards that we analyze[3]:

- **ECHO10**
- **DIF10**
- **UMM-JSON**

Collection level records (which describe a dataset) are found in all three standards. Granule level metadata (which describes a file) is currently only in the ECHO10 standard. Across all three standards, there is an average of 279 fields within a single metadata record. Until recently, these fields were being checked by hand.

1. **Date/Time fields** are checked against the W3C formats.
2. **Standard number fields** that should contain only numbers (phone contacts, geographical coordinates, etc.) are put through a check that allows only numerical values, flagging errors for symbols or letters.
3. **GCMD controlled fields** with values that are consistent with GCMD keywords are put through a hierarchy check, making sure each keyword is connected correctly.
4. **Schema controlled fields** only allow EOSDIS enumeration values. Any values outside of the appropriately controlled lists will be flagged.
5. **DOI address fields** that offer a DOI address must have a properly formatted value.
6. **Latitude/Longitude field** values must be valid coordinates, as well as self-contained; meaning the collection metadata's bounding box must contain all granules.
7. **Street address values** are not directly verified, but merely checked for their formats.
8. **Open text and URL fields** are simply checked if they have values or not. If any field contains a URL, a series of HTTPS response code checks are run to verify the health of the URL in question.

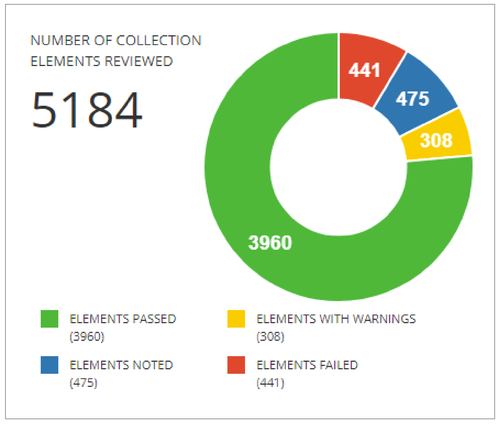
3 - It should be noted that the code does not currently evaluate the ISO metadata standard which is used for some NASA metadata. Automated checks for this standard are planned for future implementation. The ISO metadata standard has notable differences in structure, depth, and scope compared to the previously mentioned standards and therefore warrants a separate approach for automation.

Automation Success and Improvement

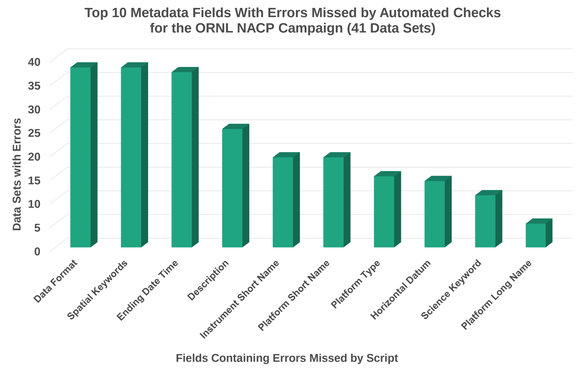
Automated checks have greatly improved the efficiency and consistency of metadata recommendations. For instance, reviewers do not have to spend time manually checking the health of URLs or validating keywords. Automated outputs also allow curators to word recommendations in a more consistent manner. The long term goal of this project is to automate the review process as much as possible. Even with automation, however, certain issues within the metadata are currently only identified via manual review. For example, the provided abstract/summary contained within the metadata may include information that is outdated, or may be lacking in important contextual information. While the scripts flag incorrect keywords, only a manual reviewer can determine whether the keywords are appropriate for the dataset. As these examples illustrate, combining both manual and automated checks allow for the highest quality metadata review.

Development and maintenance of the automated checks are ongoing. The scripts need to be updated to account for changes to the schemas, controlled vocabularies, and bugs. Scripts are also updated to include new and enhanced checks. Once the code has been modified, it is pushed to a GitHub repository where the code is automatically ingested into the dashboard.

Our code will be fully open-sourced and hosted within NASA's public GitHub upon completion of the project.



An example of metadata review metrics in the dashboard. This example includes the sum of errors flagged across 20 collection level records for a single DAAC.



The graph above illustrates the top ten metadata elements that contain errors missed by the automated checks; meaning these fields contain errors mainly found by manual reviews, even after a first-pass of automated checks.

For a sample for 41 records, 38 data sets contained errors for the *Data Format* and *Spatial Keyword* fields. The scripts flagged these elements as missing, but reviewers manually provided specific recommendations for which Data Format and Spatial Keyword should be added. This example illustrates how the automated recommendation can be improved by manual intervention. Next, with 37 missed errors, *Ending Date Time* was verified to be incorrect when checked against timestamps in the data. This error could only be identified by a manual reviewer, unlike the Data Format and Spatial Keyword fields. It is important to note that the order and name of these fields will vary depending on the campaign or DAAC.

Conclusion

- The NASA CMR Dashboard assists metadata curators in making reviews consistent and accurate.
- A combination of automated and manual reviews are still necessary at this point in order to produce the highest quality metadata.

The dashboard is an ever-changing tool, undergoing constant revisions, changes, and enhancements. In doing so, it is proving to be a tool built for posterity for the metadata community.

Contact: patrick.staton@nsstc.uah.edu

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

NASA