Metadata Deep Dive: Results from a Detailed Quality Assessment of NASA's Earth Observation Metadata

Jeanné le Roux, Kaylin Bugbee, Adam Sisco, Patrick Staton, Camille Woods, Aaron Kaulfus, Kane Cook, Jenny Wood, Rahul Ramachandran

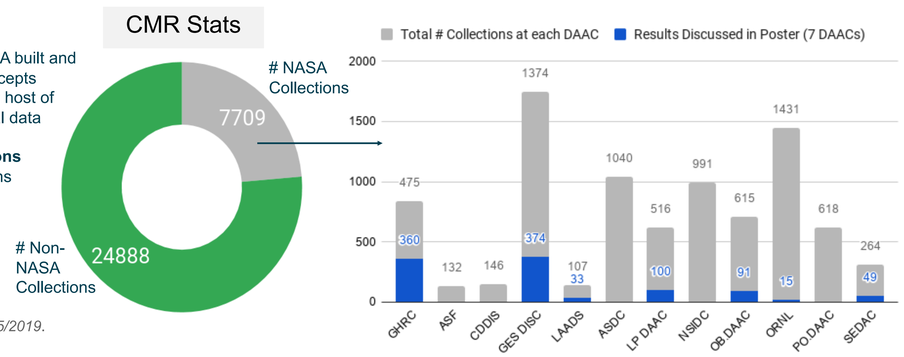
IMPACT
Interagency Implementation and Advanced Concepts Team

Introduction

- The **Common Metadata Repository (CMR)** is a high-performance repository for **Earth science metadata** records, and serves as the authoritative metadata management system for NASA's growing collection of Earth Observation data.
- Metadata records in the CMR drive the search results of the **Earthdata Search Client** (1), a web application which allows users to **search, discover, visualize, and access all of NASA's Earth observation data**.
- The quality of the metadata records in the CMR have a direct effect on the **discoverability, accessibility, and usability** of the data being described. Records that are missing information or which contain inaccurate/ outdated information create a barrier to data discovery and use.
- The **Analysis and Review of CMR (ARC)** Team at Marshall Space Flight Center has been tasked with performing **quality assessments of NASA's metadata records in the CMR**. Quality assessments involve checking metadata records for **correctness, completeness and consistency** via both automated and manual methods; documenting any findings, and providing actionable recommendations on how the metadata can be improved.
- NASA's Earth science data and metadata holdings are archived at **twelve discipline-specific data centers called Distributed Active Archive Centers (DAACs)** - findings from the assessments are packaged into reports and shared with the DAAC responsible for stewarding the metadata.
- During phase one of the project, the ARC team assessed a **subset of records from each DAAC**. Since then, each DAAC has made progress toward resolving identified issues. Once metadata updates are completed by the DAAC, the ARC team re-assesses the records to get updated metrics. **To date, ARC has collected updated metrics for a subset of records from 7/12 DAACs, the results of which are summarized in this poster.**

While the CMR is a NASA built and maintained system, it accepts metadata records from a host of national and international data partners:

- **32,597 total collections**
- 7,709 NASA collections



Numbers current as of 11/25/2019.

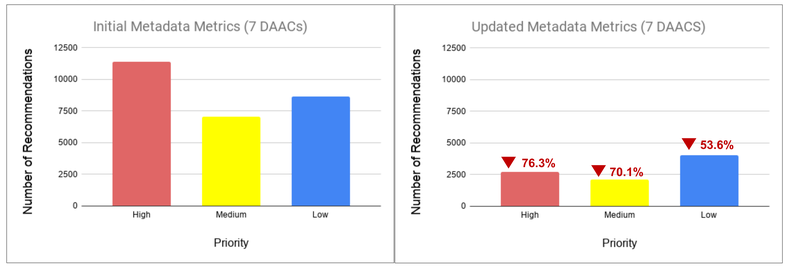
ARC Priority Matrix

- ARC has developed a **metadata quality framework in order** to assess metadata quality consistently and rigorously
- Findings are categorized based on a **priority matrix** to help prioritize issues
- The **number of high, medium, and low priority findings are used as a baseline metric** for tracking metadata improvements over time

Priority Category	Justification
Red = High Priority Issues	High priority issues are concerned with completeness, accuracy and accessibility, and are required to be addressed by the data provider. E.g. Broken data access links, outdated/incorrect information, use of invalid keywords, missing required fields
Yellow = Medium Priority Issues	Medium priority issues emphasize consistency and completeness. It is strongly encouraged that these be addressed. E.g. no link descriptions, inconsistent labeling of resources
Blue = Low Priority Issues	Low priority issues include additional information that may be provided to make the metadata more robust or complete, but likely has minimal impact to the user experience. E.g. updating links to https, providing spatial keywords
Green = No Issue	Elements flagged green are free of issues. Green flagged elements require no action on behalf of the data provider.

Results

Overall Results:



The above figure shows the total number of high, medium, and low priority findings flagged across all records after the initial assessment (left) and after updates were made by the DAACs (right). This includes a total of **1,984 records** comprised of 1,021 collection level (i.e. dataset level) records and 963 granule level (i.e. file level) records from 7 DAACs.

- During assessment, a collection record and one randomly selected granule record from the collection (if provided) are assessed for quality. At the time of initial review, 58 collections did not have accompanying file level metadata records and therefore only the collection record was assessed.

The below table summarizes the results for each DAAC:

	Red	Yellow	Blue	Total
GHRC (717 Total Records)				
GHRC Initial Metrics	6515	2459	134	9108
GHRC Updated Metrics	1072	233	405	1710
(May 2019) % Change	-83.5	-90.5	202.2	-81.2
GES DISC (748 Total Records)	Red	Yellow	Blue	Total
GES DISC Initial Metrics	1621	1862	5413	8896
GES DISC Updated Metrics	772	1018	2056	3846
(Sept. 2019) % Change	-52.4	-45.3	-62.0	-56.8
LAADS (66 Total Records)	Red	Yellow	Blue	Total
LAADS Initial Metrics	281	320	351	952
LAADS Updated Metrics	6	125	142	273
(Sept. 2019) % Change	-97.9	-60.9	-59.5	-71.3
LP DAAC (200 Total Records)	Red	Yellow	Blue	Total
LP DAAC Initial Metrics	953	818	1888	3659
LP DAAC Updated Metrics	209	263	1017	1489
(Sept. 2019) % Change	-78.1	-67.8	-46.1	-59.3
OB.DAAC (174 Total Records)	Red	Yellow	Blue	Total
OB.DAAC Initial Metrics	1210	788	434	2432
OB.DAAC Updated Metrics	536	384	314	1234
(Oct. 2019) % Change	-55.7	-51.3	-27.6	-49.3
ORNL (30 Total Records)	Red	Yellow	Blue	Total
ORNL Initial Metrics	345	159	79	583
ORNL Updated Metrics	42	23	65	130
(Aug. 2017) % Change	-87.8	-85.5	-17.7	-77.7
SEDAC (49 Total Records)	Red	Yellow	Blue	Total
SEDAC Initial Metrics	496	640	346	1482
SEDAC Updated Metrics	68	60	10	138
(Aug. 2017) % Change	-86.3	-90.6	-97.1	-90.7

Moving Targets:

Updated metrics have been shared with each DAAC. While the number of high and medium priority issues have improved across the board, they have not been eliminated completely. Often these are new issues that were not present before the record got updated. A common source of new issues are the addition of new required fields and updates to controlled vocabularies. Adoption of new metadata fields and new controlled vocabularies do not happen instantly and can take some time for the DAAC to implement. Ultimately, metadata maintenance is an ongoing task that needs to respond to changing needs over time.

Results current as of the date indicated next to % change. Some metrics may be outdated.

Iterative Process:

ARC expects that the results shown here will continue to improve over time as the results are re-iterated. This leads to the question, how many iterations of feedback are necessary? When is an assessment of a record considered to be "done"? The answer to this question is not always straight forward. At a minimum, the highest priority findings should be reconciled between the DAAC and the ARC team to ensure quality metadata in the CMR. During the next year, ARC will work to more formally define when a record is considered to be "done" going through the ARC assessment process, since the ARC project is meant to be finite in nature.

Most Common Types of Errors:

URLs	• Broken URLs • Data access URLs that do not conform to NASA requirements (ftp vs https) • No data access URLs provided at all • No URLs to essential data documentation
New Required Elements (DOIs & Collection Progress)	• DOI is a metadata concept that was recently added and is designated as required for NASA data providers • Collection State indicates the completion state of the dataset, and is also a recently added metadata element that is required • Slow adoption of new concepts by data centers explain why these fields are frequently marked red
Data Format	• Data format information not widely adopted by data centers • Not viewed as an information priority in the past, but is important to users
Abstract	• Abstracts can be particularly problematic. Common issues include: • Abstracts that are too lengthy • Non-existent • Not specific enough to describe data • Too technical for a general/global user

Discussion

- It is important to note that the number of high, medium and low priority issues is only a baseline metric which can overlook many nuances:
  - For instance, some of the issues flagged are unique from record-to-record, but in other cases, the same issue may be repeated across a large number of records. This can make a single, easily resolvable issue count toward a large number of flagged findings.
- Communication is key
  - There are always exception to the rule - sometimes there are valid reasons behind red flagged issues. For example, providing a DOI is required for all NASA datasets, so a record that is missing a DOI may be flagged red during assessment. However, there are some cases where it is not appropriate or required to provide a DOI (e.g. provisional Near Real Time datasets). These exceptions can easily be reconciled if there is communication between the DAAC and the ARC team. It is also important that these exceptions are communicated and agreed upon among the broader community.
- Despite limitations, there have been noted improvements in metadata quality throughout this effort. Processes are also being developed to help make metadata maintenance easier and to help minimize future errors.

References

(1) https://search.earthdata.nasa.gov/search

Contact: jeanne.leroux@nsstc.uah.edu

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

NASA