Eliminating Science Friction: A Metadata Quality Framework for the Earth Sciences

Jeanné le Roux[1], Kaylin Bugbee[1], Adam Sisco[1], Rahul Ramachandran[1], Patrick Staton[1], Ingrid Garcia-Solera[1], Camille Woods[1], Aaron Kaulfus[1], J.J. Miller[1], Brian Freitag[1], Peiyang Cheng[1]
(1) NASA MSFC IMPACT

IMPACT Interagency Implementation and Advanced Concepts Team

Earth Observation Data Growth

Since the launch of TIROS-1 in 1960, Earth Observation (EO) data has grown exponentially in volume. NASA alone has 32 PB of EO data (and growing) from heterogeneous sources including:

User Growth

New, easy to use software, tools, services and data formats have exposed EO data to an ever growing user base. Users can be grouped into 2 groups:

Local Users	Global Users

• Very knowledgeable about the specific scientific context within which data were collected
• Don't require much contextual information to find and use relevant data. Examples:
  • Domain Specific researchers
  • Principal investigators who originally collected the data
	• Leverage data for research and applications beyond the data's original intended use. For example:
	  • Scientists conducting research across siloed domain environments
	  • Users from the applications and decision making communities
	  • Data scientists using data in innovative new ways

Where do data and users come together?

Local users ⟫ Local data centers

Global users ⟫ Centralized, or **aggregated catalogs**

Aggregated catalogs provide a single discovery point for data from multiple sources. These catalogs bring together metadata from different data centers and presents the metadata in a unified user interface.

NASA's aggregated catalog for Earth observation data is the **Common Metadata Repository (CMR)** and the unified user interface is the **Earthdata Search** client.

Metadata in Aggregated Catalogs

Metadata sets the stage for data –

- Metadata limits & focuses attention to the relevant information about a dataset
- Metadata helps a user understand whether data is relevant to a given research problem
- Metadata makes it possible to search for data

When metadata isn't at its best, users can't –

- Find the right data
- Understand the data

When Metadata Doesn't Work…

Conducting a faceted search for 'NDVI' in Earthdata Search returns 14 datasets.

- NDVI, or the Normalized Difference Vegetation Index, is an important parameter for many applications based research questions.
  - MODIS is a key instrument for calculating NDVI, however, none of the MODIS Level 3 NDVI datasets are included in the search results. Why?
  - Due to the 'NDVI' keyword missing from the metadata

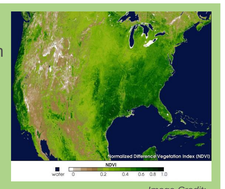
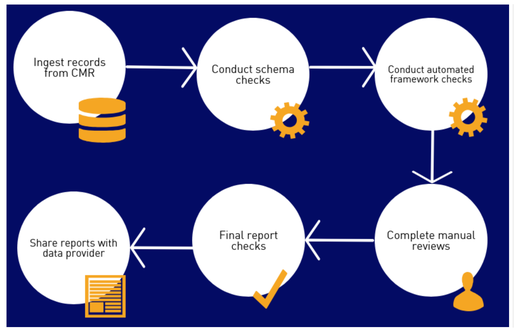


Image Credit: https://earthobservatory.nasa.gov/images/696/spring-vegetation-in-north-america

How Can We Assess Metadata Quality?

Metadata needs to be of high quality, and should be informative to both local and global users. However, finding this balance can cause metadata friction for data centers.

- NASA has established the **Analysis and Review of CMR (ARC) team** to define and assess metadata quality for EO data. The ARC team helps lower metadata friction for data centers by:
  - **Creating a metadata quality framework** to assess metadata quality consistently and rigorously
  - Leveraging automated and manual checks to assess quality
  - Building a team of reviewers with backgrounds in Earth system science, Atmospheric science, remote sensing and informatics
  - Defining a priority matrix to help prioritize issues



By leveraging this process, the ARC team has reviewed over 2,000 collection level metadata records (and a corresponding granule level record). Recommendations for improving metadata quality have been shared with all data centers.

ARC Metadata Quality Review Process

ARC Metadata Quality Framework

Quality Concept	Definition
Consistency	The extent to which metadata describes the same concepts and information in the same manner across multiple related records.
Completeness	The extent to which the metadata describes the data using all applicable metadata elements to full capacity.
Correctness or Accuracy	The extent to which the metadata reliably and correctly describes the data.

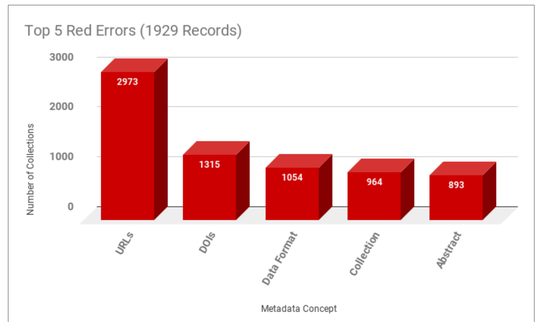
Select ARC Framework Checks

Metadata Concept	Select Automated Checks	Select Manual Checks
Temporal Information	• Temporal information adheres to ISO 8601 conventions. • Granule temporal information is within that of the parent collection.	• Temporal information in the metadata is consistent with that in the data file(s). • Temporal information has been properly translated to Coordinated Universal Time (UTC).
Data Identification	• Data are identified by a working DOI. • The responsible data center is described using GCMD conventions.	• The title is human readable and representative of the dataset. • The abstract is true to the data being described. • Identification of related journal publications describing the data.

ARC Priority Matrix

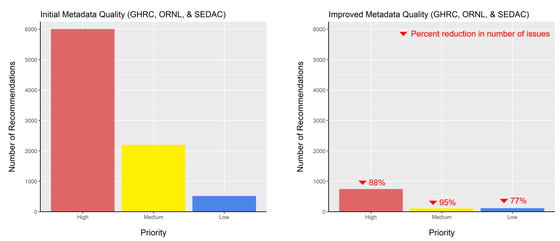
Priority Category	Justification
Red = High Priority Issues	High priority issues emphasize several characteristics of metadata quality including completeness, accuracy and accessibility. Issues flagged as red are required to be addressed by the data provider.
Yellow = Medium Priority Issues	Medium priority issues emphasize consistency and completeness. Data providers are strongly encouraged to address yellow flagged issues. If a yellow flagged issue is not addressed, the data provider will be asked to provide a justification as to why.
Blue = Low Priority Issues	Low priority issues also focus on completeness, consistency and accuracy. Any additional information that may be provided to make the metadata more robust or complete is categorized as blue.
Green = No Issue	Elements flagged green are free of issues. Green flagged elements require no action on behalf of the data provider.

Top Metadata Issues



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
DOIs & Collection Progress	• DOI is a metadata concept that was recently added and is designated as required for NASA data providers • Collection State 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 are particularly problematic. Common issues include: • Abstracts that are too lengthy • Non-existent • Not specific enough to describe data • Too technical for a global user

Metadata Improvements to Date



Combined metadata improvement metrics for 3 NASA data centers (GHRC, ORNL and SEDAC)

Lesson Learned

1. Leveraging a metadata quality framework operationally requires communication, compromise and reiteration
   - While ARC makes recommendations based on previous experience and knowledge, we are willing to compromise based on feedback from data centers
   - Therefore, ARC's metadata quality framework evolves as feedback is received and metadata standards change
2. The metadata curation process is not a "do-it-right-once-and-forget-about-it" activity and should be viewed as an iterative process
   - Data and metadata are rarely inert - scientific understanding of data evolves and changes
   - A proactive maintenance process is needed to ensure metadata is up to date, relevant and of high quality
3. Curating metadata within an aggregated catalog may require an organizational mindset change
   - Needs of global users need to be considered when curating metadata
   - Most data providers are willing to improve metadata quality as long as changes are made with sound reasoning/ guidance
   - ARC team eases this process by closing the gap for data providers between local and global needs

Conclusions

- Metadata quality can be assessed by leveraging a consistent metadata quality framework
- Metadata friction can be reduced for data centers by providing clear, easy to understand, actionable recommendations
- Improved metadata quality decreases friction for users by increasing the precision by which a dataset can be matched to a research problem
- Reducing metadata friction for data providers and scientists is still an area of opportunity

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

NASA

Contact: jeanne.leroux@nsstc.uah.edu