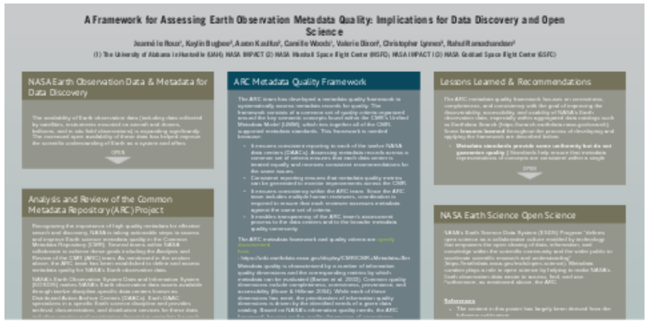
A Framework for Assessing Earth Observation Metadata Quality: Implications for Data Discovery and Open Science



Jeanné le Roux[1], Kaylin Bugbee[2], Aaron Kaulfus[2], Camille Woods[1], Valerie Dixon[3], Christopher Lynnes[3], Rahul Ramachandran[2]

(1) The University of Alabama in Huntsville (UAH); NASA IMPACT (2) NASA Marshall Space Flight Center (MSFC); NASA IMPACT | (3) NASA Goddard Space Flight Center (GSFC)

PRESENTED AT:

NASA EARTH OBSERVATION DATA & METADATA FOR DATA DISCOVERY

The availability of Earth observation data (including data collected by satellites, instruments mounted on aircraft and drones, balloons, and in situ field observations) is expanding significantly. The increased open availability of these data has helped improve the scientific understanding of Earth as a system and offers valuable data to help inform decisions, develop mitigation strategies and improve operational models.

High quality descriptive metadata is essential to enabling the effective discovery of Earth observation data to a growing number of diverse users. While metadata has many uses, descriptive metadata is important for data discovery because it limits attention to the most relevant information about a dataset. Descriptive metadata provides essential information about the data such as the title, description, keywords, the instrument used to collect the data, and the spatial and temporal extent of the data. Metadata, rather than data itself, is indexed for search in online platforms and data catalogs, making it essential for determining whether a dataset is appropriate for a given research question or application need.

Since metadata connects users to data, metadata should be as accurate and complete as possible. High quality metadata reduces the complexity of searching for data by increasing the likelihood that search terms and relationships are well matched. Conversely, poor quality metadata yields poor search results that may point users to incorrect data, or in the worst case scenario, yield no results even though suitable data exists. Once a dataset is found, incomplete or inaccurate information may impede a user's ability to determine the dataset's fitness for use for their objective. For instance, data may be inaccessible due to missing or broken data access links within the metadata, causing friction between data and data users (Edwards et al. 2011).

Given the importance of high quality, informative metadata for data discovery and use, NASA has established the Analysis and Review of the Common Metadata Repository (ARC) team to define and assess metadata quality for NASA's Earth observation data. NASA's collection of over 8000 Earth observation data products are cataloged in a centralized database called the Common Metadata Repository (CMR). This poster describes a framework developed by the ARC team to assess the quality of NASA's metadata in the CMR, lessons learned, and how this framework can be more broadly applied.

ANALYSIS AND REVIEW OF THE COMMON METADATA REPOSITORY (ARC) PROJECT

Recognizing the importance of high quality metadata for effective search and discovery, NASA is taking actionable steps to assess and improve Earth science metadata quality in the Common Metadata Repository (CMR). Several teams within NASA collaborate to achieve these goals including the Analysis and Review of the CMR (ARC) team. As mentioned in the section above, the ARC team has been established to define and assess metadata quality for NASA's Earth observation data.

NASA's Earth Observation System Data and Information System (EOSDIS) makes NASA's Earth observation data assets available through twelve discipline-specific data centers known as Distributed Active Archive Centers (DAACs). Each DAAC specializes in a specific Earth science discipline and provides archival, documentation, and distribution services for these data including creating and maintaining descriptive metadata for each data product. Each DAAC has developed unique local data architectures but is also required to conform to a common set of EOSDIS requirements that include providing metadata to the CMR. Each DAAC has its own team of metadata curators. The CMR also has personnel dedicated to maintaining the CMR database and metadata quality.

The ARC team, located within the Interagency Implementation and Advanced Concepts Team (IMPACT) at NASA's Marshall Space Flight Center, provides metadata quality assessments and improvement recommendations to each NASA data center. The ARC team serves as an independent metadata assessment group that is distinct from the twelve DAACs. The ARC team consists of Earth system scientists who have experience using the various Earth observation data types - this domain experience allows the ARC team to assess metadata within the appropriate scientific context while also considering more general users who may use NASA's Earth observation data for diverse research and application needs.

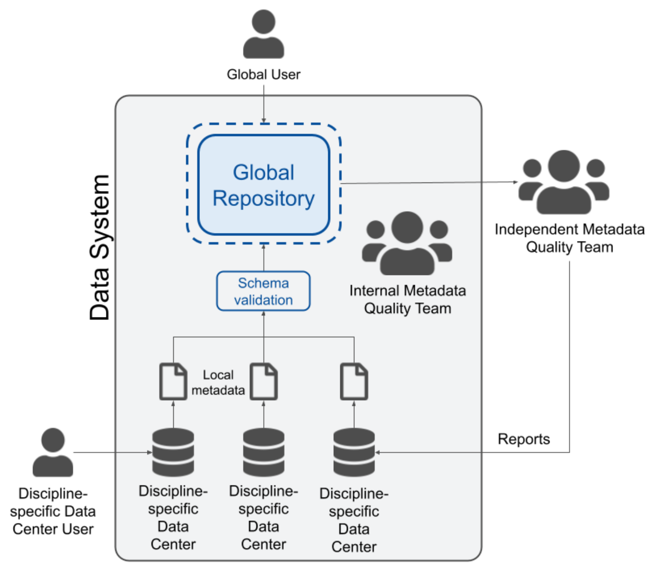


Figure 1: A conceptual model of the metadata quality assessment process within a data system (e.g. NASA EOSDIS). This data system is made up of multiple discipline-specific data centers (e.g. NASA DAACs) that contribute metadata to a centralized data catalog (e.g. the CMR). To conduct assessments, an independent quality team (e.g. ARC) systematically reviews metadata within the centralized data catalog and reports findings to the data centers (e.g. DAACs). The data center curators update the metadata and resubmit to the catalog. The discipline-specific data centers, internal metadata quality team (e.g. CMR team), and independent metadata quality team work together to improve the metadata standards and content.

ARC METADATA QUALITY FRAMEWORK

The ARC team has developed a metadata quality framework to systematically assess metadata records for quality. The framework consists of a common set of quality criteria organized around the key semantic concepts found within the CMR's Unified Metadata Model (UMM), which ties together all of the CMR-supported metadata standards. This framework is needed because:

- It ensures consistent reporting to each of the twelve NASA data centers (DAACs). Assessing metadata records across a common set of criteria ensures that each data center is treated equally and receives consistent recommendations for the same issues.
- Consistent reporting ensures that metadata quality metrics can be generated to monitor improvements across the CMR.
- It ensures consistency within the ARC team. Since the ARC team includes multiple human reviewers, coordination is required to ensure that each reviewer assesses metadata against the same set of criteria.
- It enables transparency of the ARC team's assessment process to the data centers and to the broader metadata quality community.

The ARC metadata framework and quality criteria are openly documented here (https://wiki.earthdata.nasa.gov/display/CMR/CMR+Metadata+Best+Practices%3A+Landing+Page): https://wiki.earthdata.nasa.gov/display/

Metadata quality is characterized by a number of information quality dimensions and the corresponding metrics by which metadata can be evaluated (Barton et al. 2003). Common quality dimensions include completeness, correctness, provenance, and accessibility (Bruce & Hillman 2004). While each of these dimensions has merit, the prioritization of information quality dimensions is driven by the identified needs of a given data catalog. Based on NASA's information quality needs, the ARC framework focuses on the quality dimensions of correctness, completeness, and consistency:

Correctness

- Correctness is the extent to which the metadata reliably and accurately describes the data (Bruce & Hillmann 2004; Zavalina et al. 2016). The ARC team defines metadata correctness in relation to the described data object by comparing the metadata with the actual data files and accompanying documentation. For instance, scientific variables contained within a file are compared with science keywords provided in the associated metadata record for accuracy. If a keyword in the metadata does not align with the parameters provided in the file, a reviewer recommends the keyword be removed or replaced with a more scientifically accurate one. Similarly, a reviewer may recommend additional keywords be added in order to more completely describe the scientific variables provided.

Completeness

- Completeness is the extent to which the metadata describes the data fully using all applicable metadata elements (Zavalina et al. 2016). Completeness not only measures compliance with the use of all required elements in the information model but also considers whether the metadata leverages optional elements to sufficiently describe the data (Bruce & Hillmann 2004). The ARC team assesses completeness by evaluating compliance with required elements within both the UMM and the NASA data center's local metadata standard. Additionally, recommendations are made to leverage optional elements to more completely describe the data. For example, spatial information about the data, including the horizontal datum, vertical datum, and spatial resolution, are consistently missing from many assessed metadata records. While these elements are optional, the information provided in these elements helps users assess the fitness of a dataset for a given use case. The ARC team, therefore, recommends this information be added to the metadata to provide relevant metadata to the community.

Consistency

- Consistency is the extent to which metadata describes the same semantic concepts and information in the same manner across multiple records. The ARC metadata quality framework defines consistency within the CMR and across the twelve data centers by ensuring that metadata elements are understood in a standard way (Bruce & Hillmann 2004). For example, each data center is encouraged to provide access to online resources within the metadata record. Examples of online resources include a link to the dataset landing page, the user's guide, the Algorithm Theoretical Basis Document (ATBD), available web services, and the data citation policy. Data centers are asked to ensure that similar resources are labeled consistently from record to record. For instance, if the ATBD is labeled as 'Algorithm Theoretical Basis Document (ATBD)' in one record but is labeled as 'General Documentation' in another, a user may have difficulty identifying the ATBD in the record with the more general label. Promoting consistency in how resources are labeled helps ensure a consistent experience.

The ARC metadata quality framework is used in day-to-day processes to assess NASA's metadata records within the CMR. These metadata records include collection-level (i.e. dataset-level) and granule-level (i.e. file-level) metadata. The ARC team assesses each NASA collection metadata record in the CMR and one corresponding, randomly selected granule metadata record per collection. Only one granule is assessed per collection since some collections contain millions of granules, and granule metadata is typically generated in an automated matter making it likely that an issue identified in one will be present in the other granule records from the same collection. The ARC metadata assessment process is summarized in the figure below:

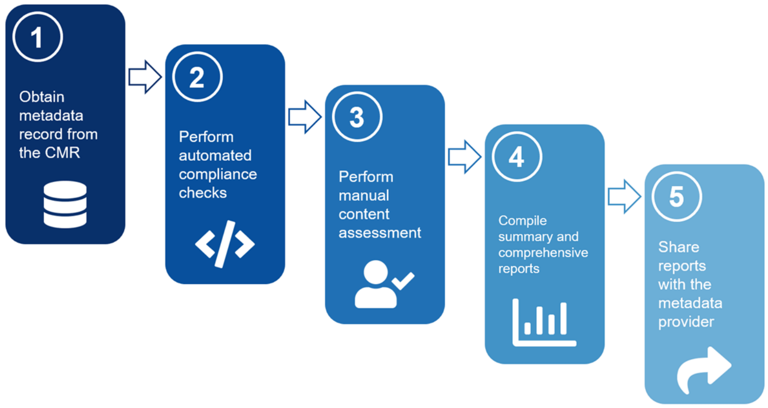


Figure 2: ARC metadata assessment process

First, a metadata record is downloaded from the CMR and a series of automated metadata checks are performed. The ARC team has developed an open-source python library called pyQuARC to process the automated checks: https://github.com/NASA-IMPACT/pyQuARC

While automated checks are effective in identifying certain issues, some errors may only be identified by a human reviewer. Thus, a manual assessment is performed by two ARC team members to identify issues and to provide actionable recommendations for improvement. Each metadata element reviewed during the assessment process is assigned a priority and a corresponding color categorization to indicate the urgency of the finding. Prioritization is provided in order to help data centers rank the findings when developing work plans and is comprised of the following categories:

- **High Priority Findings (Red):** Information that is outdated, incomplete, or objectively incorrect. For example, broken URLs, spelling or grammatical errors, or absence of required information. High-priority findings typically address barriers to data accessibility or use and are therefore required to be resolved by the data provider.
- **Medium Priority Findings (Yellow):** Highly recommended suggestions that place an emphasis on consistency and completeness. Data providers are strongly encouraged to address medium-priority issues and are encouraged to provide a rationale for any findings not addressed.
- **Low Priority Findings (Blue):** Typically focus on minor inconsistencies or missing information that may make the metadata more robust. Low-priority findings are unlikely to have any significant impact on data discoverability.
- Elements with no issues are flagged as **green** to indicate that the element was reviewed by the ARC team and no findings were identified.

A priority matrix has been developed by the ARC team for each metadata concept within the UMM. The content in each metadata element may be assigned a different priority based on how the finding affects the three metadata quality dimensions of correctness, completeness, or consistency. For example, findings about URLs may be assigned a high, medium, or low priority depending on the significance of the finding. The ARC team has documented the priority matrix for each metadata element and has made the information publically available (https://wiki.earthdata.nasa.gov/display/CMR/CMR+Metadata+Best+Practices%3A+Landing+Page).

Upon completing all automated and manual assessments for a data center, the findings are compiled in reports and shared directly with the data center. Upon receiving the ARC assessment reports, each data center formulates a work plan for addressing the report findings. Once updates are complete, the data center notifies ARC that the metadata is ready for reassessment. Metadata elements that are updated per the ARC recommendations are marked as resolved while elements that have not been updated and any new issues are reported back to the data center. This process iterates until the data center works off all reported high-priority findings or negotiates an agreement with ARC for unaddressed issues.

Thus far, all twelve data centers have received ARC reports on assessed metadata records and are either in the process of updating metadata or have completed all requested updates. A subset of records from nine data centers has been re-evaluated by ARC thus far. Upon reassessment of the nine data centers, the number of findings identified after updates resulted in substantial improvements of 71%, 60%, and 30% in high, medium, and low priority findings, respectively. The remaining findings have been reported to each data center and results are expected to continue to improve with subsequent iterations of the process. Iterations will continue until all high-priority findings have been addressed.

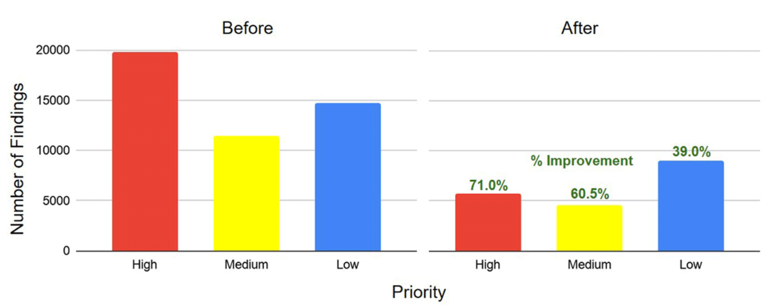


Figure 3: The cumulative number of high (red), medium (yellow), and low (blue) priority findings for the nine data centers upon initial assessment (left) and after reassessment (right).

These initial results demonstrate that descriptive metadata can be assessed and improved through the consistent application of a metadata quality framework and through close community collaboration and communication. While the specifics of the framework and quality criteria have been tailored for the NASA Earth observation use case, the principles and basic process have the potential to be adapted and reused for other data catalogs and science disciplines.

LESSONS LEARNED & RECOMMENDATIONS

The ARC metadata quality framework focuses on correctness, completeness, and consistency with the goal of improving the discoverability, accessibility, and usability of NASA's Earth observation data, especially within aggregated data catalogs such as Earthdata Search (https://search.earthdata.nasa.gov/search). Some **lessons learned** throughout the process of developing and applying the framework are described below:

- **Metadata standards provide some uniformity but do not guarantee quality |** Standards help ensure that metadata representations of concepts are consistent within a single standard, but do not guarantee correctness or completeness (Stvilia, Gasser & Twidal 2004). For example, the initial ARC assessments showed that most metadata records only use about 40% of the available elements in a standard. These records may meet minimum syntactic requirements, but do not utilize all available and applicable concepts that may provide greater context for a user. Even when minimum requirements are met, the quality of content provided can be improved, as shown by the numerous quality findings. Well constructed, compliant, and easy-to-use metadata authoring tools can help ensure a minimum level of metadata quality is met for a single standard. Metadata written using specialized tools such as NASA's Metadata Management Tool (MMT) requires that all metadata content entered using the tool meet minimum requirements. Nonetheless, metadata can still be incorrect and incomplete even when using a tool - until tools use new techniques to check metadata for scientific correctness, there is always a chance that there resulting metadata can be incorrect.
- **Metadata quality and consistency improve when metadata standards are inclusive of heterogeneous data types |** NASA's Earth observation metadata standards are largely designed to describe homogeneous, high volume, well defined data from traditional satellite missions (Parsons & Fox 2013). However, as NASA's data holdings have expanded to include heterogeneous data sources, these metadata models have limited the description of an array of diverse data including observations from satellites with specialized instruments such as synthetic aperture radar (SAR) and lidar, data from smallsat constellations, and airborne and in situ field observations. For example, airborne field investigation metadata needs are not well met within the existing metadata standards. Airborne and field data are organized around contextual campaign information that users want to use to discover data, but lack of support and flexibility in existing metadata models have led some of the airborne investigation data centers to include this information in the metadata using inconsistent and ad hoc approaches. Groups like the ARC team and the Airborne Data Management Group (https://earthdata.nasa.gov/esds/impact/admg) are working with the data centers to help build consistent metadata quality approaches until the standards evolve to meet heterogeneous data needs. Longer term, the addition of new metadata elements designed to help serve the search and discovery needs of these specific user communities would help ensure consistency across the CMR.
- **High level data management principles and metadata quality frameworks are only effective if actionable recommendations are provided |** Based on ARC's experience, the effectiveness of high level, domain independent data management principles and metadata quality frameworks are limited at best in the absence of actionable recommendations. Principles and recommendations from high level frameworks, such as the FAIR data principles (Wilkinson et al. 2016), are deliberately open ended and are only meant to guide the unique implementation choices for each archive/data center. This flexibility may be beneficial if each data center is considered in isolation. However, the reality is that most data centers are part of a larger ecosystem, with data and metadata being shared to any number of aggregated data catalogs. When each data center interprets broad principles for individual needs, interoperability and understanding for a broader user community may be sacrificed. For example, one FAIR data principle recommends that "data are described with rich metadata." While this is a good recommendation, it is still too broad and ambiguous for day-to-day implementation of authoring and maintaining high quality metadata. In the ARC team's experience, data centers want to provide high quality metadata that meet both discipline-specific and broader community needs, however, independently determining what rich, high quality metadata looks like has resulted in twelve unique interpretations. The ARC team's definitive and independent recommendations, along with the priority classification of those recommendations, help provide more concrete guidance.
- **High quality metadata does not guarantee consistent data accessibility |** High quality metadata increases the discoverability of data and also increases the likelihood that data will be accessible and usable, but does not guarantee consistency in how data access is implemented. For example, data access methods across NASA's twelve data centers are variable, making a consistent data access experience difficult. Across the twelve data centers, there are more than 61 data tools available for searching, ordering, subsetting, filtering, projection, geolocation, and visualization (Liu et al. 2020). Some data centers rely entirely on these ordering tools for access while others provide direct file access. For those data centers that provide direct file access, the file access structures vary, making it challenging for a user to easily find correct data. Metadata may contain all the required information needed to access data but the variety of data access pathways along with the learning curve associated with specialized data access tools deters from a consistent data access experience. However, with the transition to data in the cloud, there have been efforts such as NASA's Harmony project (https://harmony.earthdata.nasa.gov/) to streamline and make seamless the ability to analyze Earth observation data from different data centers. With such efforts, the consistency of the user experience in interacting with data is expected to improve.
- **The metadata aggregator's role should be expanded to both reinforce metadata quality and to enable broader data discovery |** To date, many metadata aggregators, including the CMR, assume the original local record from the data center is the source of truth for metadata. While some aggregators may apply rules that convert values into human-readable text within the system, most metadata quality issues must be fixed at the source. Updated records must then be pushed from the local database to the aggregated catalog. We suggest that the aggregator role be expanded to feel empowered to make changes to metadata in the aggregated environment, as long as those changes are communicated back to the local data center/source database. The aggregator may

consider making two types of metadata changes: transformation or augmentation changes. Transformation changes modify metadata based on the information already provided in the record (Hillman 2008). These changes include resolving typographical errors, removing deprecated elements, and detecting duplicate records. Transformation changes are designed to only correct existing metadata, and should not modify or replace existing metadata that is already correct. The semantics of the record should be kept in tact. Augmentation changes, on the other hand, leverage the information provided in the metadata to add new information or value to the aggregated record (Hillman 2008). Examples of such additions include detecting and adding the data format to the metadata, using machine learning techniques to add scientific keywords to the metadata, and gathering and leveraging collection statistics (Stvilia, Gasser & Twidal 2004). To ensure metadata quality is maintained cross-enterprise, the aggregator may consider using automated techniques to communicate quality changes back to the stewards of the source metadata.

Recommendations

In an era of exponential data volume growth and broader data use beyond domain specific science communities, high quality metadata is critical for both discovering data and understanding the scientific context of data. The ARC project has demonstrated that metadata quality can be assessed and improved through the consistent application of a metadata quality framework and through close community collaboration and communication. Based on ARC's experience, the metadata improvement process achieves the best results when both automated and manual evaluation methods are used that are systematic and transparent to all involved stakeholders.

Providing an independent perspective to the metadata quality assessment process is also beneficial. ARC's position as neutral reviewers benefitted the overall process by assessing metadata from an enterprise, data system perspective instead of an archive-by-archive approach. While most data centers excel at considering the needs of their user community, the needs of more general users may not always be prioritized due to the ambiguity surrounding those needs. The ARC team process reduces ambiguity by providing easy to understand and actionable recommendations.

Data systems who wish to implement a similar metadata quality assessment process should consider the following recommendations:

- First, establish metadata quality priorities, such as discoverability, accessibility, or provenance, as early as possible in the process
- Be as transparent as possible with all stakeholders as early as possible in the process, including providing open access to detailed metadata quality assessment checks and openness to stakeholder feedback
- Be as flexible and adaptable as possible when creating and maintaining quality assessment checks, as this process is iterative and will evolve with stakeholder feedback
- Recognize that time and resources may be limited - providing prioritized recommendations helps all stakeholders determine which findings to address first
- Make any resources available for reuse by the community if possible. For instance, the ARC team documentation is openly available (https://wiki.earthdata.nasa.gov/display/CMR/CMR+Metadata+Best+Practices%3A+Landing+Page). Also open-sourced is the ARC team's metadata quality review tool (https://github.com/nasa/cmr-metadata-review) and suite of automated metadata quality checks (https://github.com/NASA-IMPACT/pyQuARC)

NASA EARTH SCIENCE OPEN SCIENCE

NASA's Earth Science Data System (ESDS) Program "defines open science as a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding" (https://earthdata.nasa.gov/esds/open-science). Metadata curation plays a role in open science by helping to make NASA's Earth observation data easier to access, find, and use. Furthermore, as mentioned above, the ARC project strives to make its processes for evaluating metadata as transparent as possible by providing the following resources. These resources are linked in the list and cited in the references:

- Publication describing ARC's metadata quality framework (http://doi.org/10.5334/dsj-2021-017)
- ARC Metadata Quality Framework quality criteria and documentation (https://wiki.earthdata.nasa.gov/display/CMR/CMR+Metadata+Best+Practices%3A+Landing+Page)
- Open source code for the CMR Metadata Curation Dashboard (https://github.com/nasa/cmr-metadata-review) tool, used to assess and generate metadata quality reports
- Open source code for pyQuARC (http://https://github.com/NASA-IMPACT/pyQuARC), a python library used to process ARC's automated metadata quality checks. pyQuARC can be customized to add new checks, modify existing checks, and to support different metadata standards.

References

- The content in this poster has largely been derived from the following publication:

Bugbee, K., le Roux, J., Sisco, A., Kaulfus, A., Staton, P., Woods, C., Dixon, V., Lynnes, C. and Ramachandran, R., 2021. Improving Discovery and Use of NASA's Earth Observation Data Through Metadata Quality Assessments. ***Data Science Journal*****, 20(1), p.17. DOI: http://doi.org/10.5334/dsj-2021-017 (http://doi.org/10.5334/dsj-2021-017)**

- Wiki space containing CMR Metadata Best Practices and ARC Metadata Quality Framework documentation (i.e. ARC Priority Matrix quality criteria for each UMM metadata concept): https://wiki.earthdata.nasa.gov/display/CMR/CMR+Metadata+Best+Practices%3A+Landing+Page

- pyQuARC: Open Source Library for Earth Observation Metadata Quality Assessment: https://github.com/NASA-IMPACT/pyQuARC

- CMR Metadata Curation Tool (used for ARC metadata quality reporting): https://github.com/nasa/cmr-metadata-review

- Airborne Data Management Group: https://earthdata.nasa.gov/esds/impact/admg

- NASA Earthdata Harmony project: https://harmony.earthdata.nasa.gov/

- Edwards, PN, et al. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Final report of the workshop "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures." Available at https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf?sequence=3&isAllowed=y (https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf?sequence=3&isAllowed=y)

- Barton, J, Currier, S and Hey, JMN. 2003. Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice. In: DC-2003--Seattle Proceedings. Seattle, WA, 39–48. Available at https://dcpapers.dublincore.org/pubs/article/viewFile/732/728 (https://dcpapers.dublincore.org/pubs/article/viewFile/732/728).

- Bruce, TR and Hillmann, DI. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: *Metadata in Practice*. Cornell University Library: ALA Editions. Available at https://hdl.handle.net/1813/7895 (https://hdl.handle.net/1813/7895).

- Zavalina, OL, et al. 2016. Developing an empirically-based framework of metadata change and exploring relation between metadata change and metadata quality in MARC library metadata. *Procedia Computer Science*, 99: 50–63. DOI: https://doi.org/10.1016/j.procs.2016.09.100 (https://doi.org/10.1016/j.procs.2016.09.100)

- Stvilia, B, Gasser, L and Twidale, MB. 2004. Metadata Quality for Federated Collections. In: Al-Hakim, L (ed.), *Challenges of Managing Information Quality in Service Organizations*. IGI Global. pp. 154–186. DOI: https://doi.org/10.4018/978-1-59904-420-0.ch008 (https://doi.org/10.4018/978-1-59904-420-0.ch008)

- Parsons, MA and Fox, PA. 2013. Is Data Publication the Right Metaphor? *Data Science Journal*, 12: WDS32–WDS46. DOI: https://doi.org/10.2481/dsj.WDS-042 (https://doi.org/10.2481/dsj.WDS-042)

- Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3. DOI: https://doi.org/10.1038/sdata.2016.18 (https://doi.org/10.1038/sdata.2016.18)

- Liu, Z, et al. 2020. Creating Data Tool Kits That Everyone Can Use. *Eos*, 101. DOI: https://doi.org/10.1029/2020EO143953 (https://doi.org/10.1029/2020EO143953)
- Hillmann, DI. 2008. Metadata Quality: From Evaluation to Augmentation. *Cataloging & Classification Quarterly*, 46(1): 65–80. DOI: https://doi.org/10.1080/01639370802183008 (https://doi.org/10.1080/01639370802183008)

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program.

ABSTRACT

The Common Metadata Repository (CMR) contains metadata records describing NASA's collection of over 8,000 Earth observation data products. The Analysis and Review of CMR (ARC) Team at Marshall Space Flight Center assesses the quality of these metadata records. Metadata, rather than the data itself, is indexed for search in both discipline-specific data centers and global or aggregated catalogs (such as Earthdata Search), making it essential for determining whether a data product is appropriate for a given research question or application need. Since metadata connects users to data, it should be as accurate and complete as possible in addition to meeting minimum database requirements. The ARC team has developed a metadata quality framework by which to assess quality. The framework consists of a set of quality criteria that converge around the dimensions of correctness, completeness, and consistency, with the goal of improving the discoverability, accessibility, and usability of NASA's Earth Observation data. The application of the framework has resulted in a measurable improvement in NASA's metadata quality. Key aspects of the framework's success are the ability to systematically evaluate metadata and provide actionable quality improvement recommendations. Lessons learned from the project will be shared along with implementation details which may be relevant to other science disciplines. By aiming to make data more discoverable and accessible to a broad user community, the ARC metadata quality framework helps contribute to NASA's commitment to open science.