Improving Discovery and Use of NASA's Earth Observation Data Through Metadata Quality Assessments

RESEARCH PAPER

KAYLIN BUGBEE
JEANNÉ LE ROUX
ADAM SISCO
AARON KAULFUS
PATRICK STATON
CAMILLE WOODS
VALERIE DIXON
CHRISTOPHER LYNNES
RAHUL RAMACHANDRAN

Author affiliations can be found in the back matter of this article

ABSTRACT

High quality descriptive metadata is essential to enabling the effective discovery of Earth observation data to a growing number of diverse users. In this paper, we define a framework to assess the quality of NASA's Earth observation metadata with the overarching goal of improving the discoverability, accessibility and usability of the data it describes. The framework, developed by the Analysis and Review of the Common Metadata Repository (ARC) team, focuses on the metadata quality dimensions of correctness, completeness, and consistency. The methodology used by the ARC team to implement the framework is described, as well as best practices, lessons learned and recommendations for implementing similar metadata quality assessment processes. Initial results from the project indicate that this methodology, in combination with community and stakeholder collaboration, is effective in improving metadata quality.

CORRESPONDING AUTHOR:
Kaylin Bugbee
NASA, US
kaylin.m.bugbee@nasa.gov

KEYWORDS:
Metadata quality; Stewardship; Accessibility; Usability; Information Management; Assessment Framework

TO CITE THIS ARTICLE:
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: 17, pp. 1–15. DOI: *https://doi.org/10.5334/dsj-2021-017*

1. INTRODUCTION

Since the launch of the TIROS-1 weather satellite in 1960, the availability of Earth observation data has expanded significantly. The increased open availability of these data has helped improve the scientific understanding of Earth as a system and has also transformed 'environmental management, decision making and operational modeling environments' by providing data to inform decisions, develop mitigation strategies and improve operational models and activities (Brown et al. 2013; Overpeck et al. 2011). Additionally, Earth observation data are often 'found to be useful for additional purposes not foreseen during the development of the observation system' (OSTP 2016). Novel uses of these data, in combination with the development of easy to use software, tools, services and data formats, have exposed Earth observation data to a growing audience.

Earth observation data is primarily discovered through two mechanisms: discipline-specific data centers and global catalogs (Edwards et al. 2007). Discipline-specific data centers typically serve a specific scientific community and provide pertinent information and services needed by the community. Discipline-specific data center users are knowledgeable about the scientific context within which the data were collected and are generally familiar with the information and services provided by the data center. Discipline-specific data center users include domain specific research scientists and principal investigators who originally collected the data. A prototypical example of a discipline-specific data center is the Alaska Satellite Facility Distributed Active Archive Center (*ASF DAAC*). The ASF DAAC serves as NASA's archive of synthetic aperture radar (SAR) data from a variety of satellites and aircraft sources. The ASF DAAC's search and discovery tools assume some knowledge of SAR and SAR observing platforms. While these tools may work well for SAR subject matter experts, the ease of use may not easily translate to users outside of the community. Examples of other discipline-specific archives include the World Data Center for Climate/CERA at DKRZ (*WDCC*), the Crustal Dynamics Data Information System (*CDDIS*) and the *Cambridge Crystallographic Data Centre*.

Global catalogs, on the other hand, aggregate, or link together, data from discipline-specific data centers into a centralized location. Global catalogs expand the reach of data by exposing data to a broader community of users via a single consolidated environment. Global catalog users seek data for research and applications beyond the data's original intended use. Global users include scientists conducting interdisciplinary research, users from the applications and decision-making communities and data scientists using data in innovative ways. Examples of global catalogs include the United States government's open data portal (*Data.gov*), NASA's *Earthdata Search* (Liu et al. 2020) and the European Space Agency's Federated Earth Observation (*FedEO*) portal. For NASA's Earth Science Data System, the Earthdata Search, supported by the Common Metadata Repository (CMR), is the global catalog while the twelve NASA distributed active archive centers (DAACs) are the discipline-specific data centers.

Descriptive metadata is the key mechanism that facilitates data discovery within both discipline-specific data centers and global catalogs. While metadata has many uses, descriptive metadata is important for data discovery because it limits or focuses attention to the most relevant information about a dataset. Descriptive metadata provides essential information about the data such as the title, abstract, keywords, the instrument used to collect the data and the geographic and temporal extent of the data. Metadata, rather than the data itself, is indexed for search in both discipline-specific data centers and global 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. Searching for relevant data is a complex task that requires 'the articulation of an information need, often ambiguous, into precise words and relationships that match the structure of the system being searched' (Borgman 1986). High quality metadata reduces the complexity of searching for data by increasing the likelihood that search terms and relationships are well matched. However, 'if metadata quality is poor so is the discovery… of objects' (Tani, Candela & Castelli 2013), resulting in poor search results that may point the user to incorrect data. Poor quality metadata may even mean that 'a resource is essentially invisible within a repository or archive and remains unused' (Barton, Currier & Hey 2003).

Once a dataset is found, incomplete or inaccurate metadata may further deter users from accessing or using the associated data. Incorrect or missing information may impede a user from determining the dataset's fitness for a particular research question or new application. Similarly, descriptive metadata that is sufficient for a discipline-specific data center's user community may not be as effective in a global catalog (Barton, Currier & Hey 2003) where local, contextual information cannot be assumed (Tani, Candela & Castelli 2013). In the worst case scenario, data is inaccessible due to broken or missing data access links within the metadata. These information gaps contribute to friction (Edwards et al. 2011) between data and data users, reducing the likelihood that data will be discovered and used for novel research and applications.

Recognizing the importance of high quality, informative metadata for both discipline-specific data centers and global catalogs, NASA has established the Analysis and Review of CMR (ARC) team to define and assess metadata quality for Earth observation data and to lower friction for both metadata providers and data users. The ARC team serves as a social gateway (Edwards et al. 2007) between the various communities of practice across NASA by providing common quality recommendations that work within existing systems for both discipline-specific data centers and the global catalog. The goal of this paper is to encapsulate the ARC metadata quality assessment process, along with best practices and lessons learned from this project, so that other agencies and organizations with global catalogs may benefit. In this paper, we present a quality framework for assessing NASA's Earth observation metadata, a methodology for implementing the framework in an actionable manner, and a process for collaborating with metadata authors to improve quality. Initial results of the project are presented along with lessons learned from the ARC team's efforts to date. We conclude with recommendations for implementing a similar metadata quality assessment process along with our vision for the future of metadata quality assessment.

2. NASA'S METADATA INFRASTRUCTURE AND TEAMS

2.1 NASA'S METADATA INFRASTRUCTURE

NASA's Earth Science Division (ESD) seeks to better understand Earth as a system by collecting Earth observations from satellites, aircraft, balloons, field measurements and model outputs. The datasets created by these observations are freely and openly available to a wide community of users. NASA's Earth Observing System Data and Information System (EOSDIS) makes these data available through twelve discipline-specific data centers known as Distributed Active Archive Centers (DAACs). Each data center specializes in a specific scientific discipline and provides archival, documentation and distribution services for these data including creating and maintaining descriptive metadata for each dataset. Each data center has developed unique local data architectures (NASA 2017) but are also required to conform to a common set of EOSDIS requirements that include providing metadata to NASA's Common Metadata Repository (CMR). The CMR is an aggregated catalog that serves as the foundation for EOSDIS's global discovery interface, Earthdata Search, which enables data search, comparison, visualization and access across all EOSDIS data holdings.

In order to serve each data center's various metadata needs, the CMR supports the ingestion of metadata in five different metadata standards. These five standards include the Directory Interchange Format (DIF), the Earth Observing System Clearinghouse Format (ECHO10), the International Organization for Standardization (ISO)'s standards ISO 19115-1 and 19115-2, and the Unified Metadata Model (UMM). Metadata interoperability between these standards is made possible within the CMR by the UMM, which serves as both a metadata model and as a crosswalk between the supported metadata standards. The UMM crosswalk lowers barriers to providing metadata to the CMR for the discipline-specific data centers and increases interoperability within the CMR. In the end, assembling metadata from these various organizations and standards into a single, common repository vastly improves the discovery, access and use of disparate Earth observation data collections (Baynes & Mitchell 2017) for all users.

2.2 NASA'S METADATA QUALITY TEAMS

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 CMR.

Several teams within NASA collaborate together to achieve these goals including the Analysis and Review of CMR (ARC) team, the discipline-specific data center metadata curators and the EOSDIS Evolution and Development (EED2) metadata quality team (***Figure 1***). 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 EOSDIS data centers. The ARC team consists of Earth system scientists who have experience using the various Earth observation data types and associated tools used to analyze these data. This domain experience allows the ARC team to assess metadata within the appropriate scientific context and also consider the needs of global users who may use the data for diverse research needs.

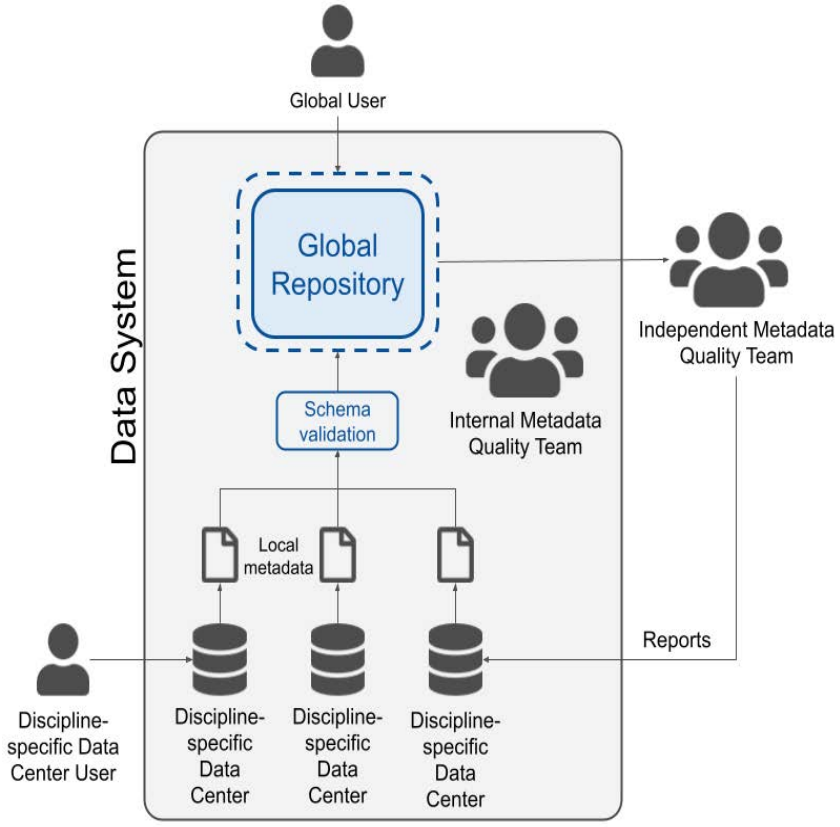

Figure 1 A conceptual model of the metadata quality assessment process within a data system. A data system is made up of discipline-specific centers that contribute metadata to a centralized global catalog. To conduct assessments, an independent quality team systematically reviews metadata within the global catalog and reports findings to the discipline-specific data centers. The discipline-specific data center curators update the metadata and resubmit it to the global catalog, improving the quality. The discipline-specific data centers, the internal metadata quality team and the independent quality team work together to improve the metadata standards and content. For NASA, EOSDIS is the data system, the DAACs are the discipline-specific data centers, EED2 is the internal metadata quality team and ARC is the independent metadata quality team.

The ARC team collaborates closely with both the discipline-specific data center metadata curators and the EED2 team to support CMR metadata quality. First, the ARC team works closely with the metadata curators at each NASA data center to improve metadata quality in the local database that is in turn provided to the CMR. Second, the ARC team collaborates with the EED2 team, the managers of the UMM governance process, to provide recommendations on needed UMM, CMR and keyword changes.

3. METADATA QUALITY ASSESSMENT METHODOLOGY

3.1 THE ARC METADATA QUALITY FRAMEWORK

The ARC team has created a metadata quality framework (CMR Metadata Best Practices: Landing Page, 2020) to systematically assess metadata records for quality. The framework consists of a common set of assessment criteria organized around the key semantic concepts found within the UMM. This framework is needed for several reasons. First, the framework ensures consistent reporting to each of the twelve NASA data centers. 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. This level of consistency also ensures that metadata quality metrics can be generated to monitor improvements across the CMR. Second, the framework 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. Lastly, the framework enables transparency of the ARC team's assessment process to the data centers and to the broader metadata quality community. The ARC metadata quality framework is openly documented (CMR Metadata Best Practices: Landing Page, 2020) so that the data centers can understand both the criteria used to assess the metadata quality and to understand the reports provided by the ARC team. This open documentation enables both the NASA data centers and the broader community to discuss and leverage the framework practices.

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, consistency, timeliness and accessibility (Bruce & Hillmann 2004). While each of these dimensions have merit, the prioritization of the dimensions are driven by the identified needs of a given metadata catalog. The prioritization of information quality dimensions for the ARC framework was built upon lessons learned during the Climate Data Initiative (CDI) curation effort, which brought together federal climate-relevant data in a global catalog, *Data.gov*/climate, to make climate data more accessible to a broad user community (Ramachandran et al. 2016). The CDI metadata quality framework was primarily limited to assessing metadata correctness, with a secondary focus on data accessibility. Based on NASA's information quality needs, the ARC framework expanded the CDI framework to focus not only on correctness but also on completeness and consistency to support discoverability in both the discipline-specific data centers and the global catalogs. The ARC framework defines these key metadata quality dimensions as follows:

- *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 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. Lastly, the ARC team assesses the UMM itself for completeness and provides feedback on concepts that are missing from the information model. For example, one concept that was missing from the UMM was data format, which is often critical information for global catalog users to determine whether a dataset is usable. Based on this identified user need, the ARC team's recommendation to include data format information in the UMM has been adopted.
- *Consistency*
  Consistency is the extent to which metadata describes the same semantic concepts and information in the same manner across multiple records. Since metadata in the CMR is maintained by twelve data centers using five different metadata standards, there is a need to decrease the variance across the CMR. Decreasing the variance reduces the number of false positives or false negatives returned when searching for data. Variance reduction also presents a more cohesive experience for users and makes it

easier for users to compare data products from different data centers in an aggregated environment. 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) and attempts to 'increase the value of a metadata object for the non-local users...without decreasing its value to local users' (Stvilia et al. 2007). 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 for users exploring online resources within the Earthdata Search client.

3.2 METADATA QUALITY ASSESSMENT PROCESS

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 metadata which describe 'an entire set of data products or files' and granule level metadata which describe 'a single instance (granule) within a data collection' (Khalsa et al. 2011). The ARC team assesses each 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. Granule metadata is typically generated in an automated manner, making it likely that an issue identified in one granule record will be present in the other granule records from the same collection.

The assessment process begins by selecting a collection level metadata record for review. The collection record is downloaded in the discipline-specific data center's metadata standard using the CMR API (***Figure 2***, step 1). Once the record is downloaded, a series of automated metadata quality checks are performed (***Figure 2***, step 2). These automated checks leverage both syntactic and semantic checks in order to 'enable humans to use their time to make more sophisticated assessments' (Bruce & Hillmann 2004). Examples of automated checks are described in ***Table 1***. While automated checks are effective in identifying certain issues, some errors may only be identified by a human reviewer. For example, a human reviewer is needed to assess whether the abstract accurately describes the data in an understandable manner or whether a URL points as directly as possible to the correct data (***Table 1***). Thus, a manual assessment is performed by two ARC team members to identify issues and to provide actionable recommendations for improvement (***Figure 2***, step 3). The manual assessment also considers the metadata record as a cohesive unit and places an emphasis on whether the record conveys information that is helpful to both the discipline-specific data center users and global catalog users.

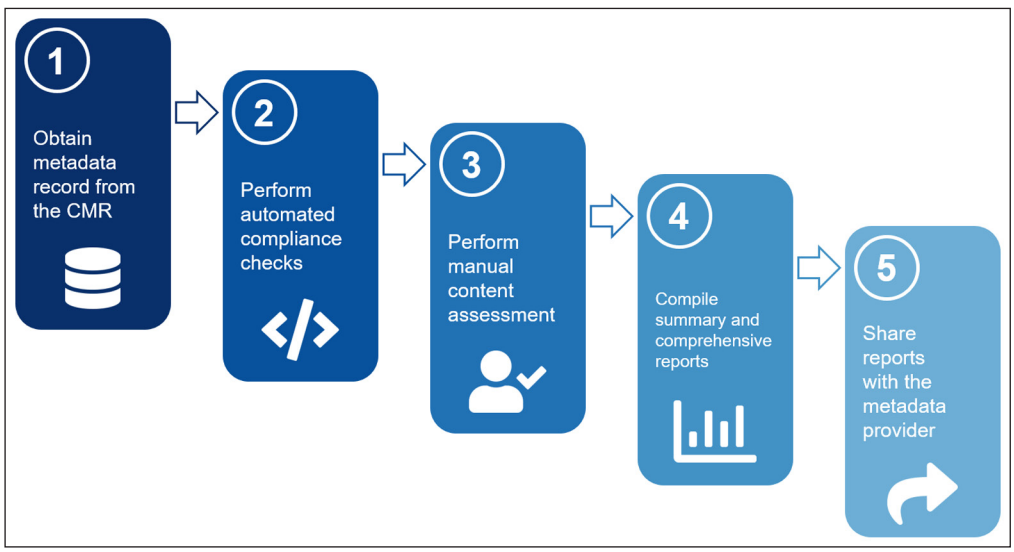

Figure 2 The ARC metadata assessment process.

	AUTOMATED CHECKS	MANUAL CHECKS
Data Identification	• Data are identified by a functioning unique identifier (e.g. DOI). • The responsible data center is identified using a controlled keyword list.	• The title is human readable and representative of the dataset. • The abstract accurately describes the data. • Key journal publications describing the data are included.
Descriptive Keywords	• Descriptive science keywords conform to GCMD conventions and/or ISO 19115 topic categories.	• The science keywords accurately describe the data to which they are applied.
URLs	• URLs are responsive and do not redirect. • FTP protocol is not utilized.	• Data access URLs point as directly to the data as possible. • Only links to relevant online resources are included.
Acquisition Information	• Earth observation platform and instrument names conform to GCMD conventions.	• Reported data collection was during a time when the acquiring instrument was active.

Table 1 Select automated and manual checks performed by the ARC team during the assessment process.

Each metadata element evaluated during the assessment process is assigned a priority and a corresponding color categorization to indicate the urgency of the identified finding (***Table 2***). Prioritization is provided in order to help the discipline-specific data centers rank the findings when developing work plans. High priority findings, which are flagged as red, focus on information that is outdated, incomplete, or objectively incorrect. For example, broken URLs, spelling or grammatical errors or an absence of required information are all classified as high priority findings. High priority findings typically address barriers to data accessibility or use and are therefore required to be resolved by the metadata provider. Medium priority findings, which are flagged as yellow, are 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 that are not addressed. Low priority findings typically focus on minor inconsistencies or missing information that may make the metadata more robust. Low priority findings, which are flagged as blue, are unlikely to have any significant impact on data discoverability and are included in ARC's reports for completeness. Lastly, elements with no issues are flagged as green to indicate that the element was reviewed by the ARC team and no findings were identified.

PRIORITY CATEGORIZATION	JUSTIFICATION
Red = High Priority Findings	Emphasizes metadata completeness, accuracy and data accessibility. Metadata that fails to meet CMR requirements or that are factually incorrect constitute a high priority finding. Examples: • Broken or missing data access URL • Non-compliance to controlled vocabulary Metadata fields flagged as red are required to be addressed by the data center.
Yellow = Medium Priority Findings	Emphasizes metadata completeness and consistency - recommendations focus on ways to help improve data discoverability and usability that go beyond CMR requirements. Examples: • A URL is missing a description. While not required, descriptions provide important context for the URL. • The same resource is labelled differently between metadata records Data centers are highly encouraged to address yellow findings and are encouraged to provide a rationale for unaddressed items.
Blue = Low Priority Findings	Documents minor metadata consistency, completeness and accuracy issues. Examples: • URLs that need to be updated from the 'http' to 'https' protocol • A DOI is provided but the DOI Authority is not specified Addressing blue findings are optional and up to the discretion of the data center.
Green = No Findings/Issues	Metadata elements flagged green are free of issues and require no action on behalf of the data center.

Table 2 The ARC team's assessment priority matrix. A priority matrix is documented for each metadata concept and identifies the criteria that indicate whether a finding should be flagged as high, medium or low priority.

A priority matrix has been developed by the ARC team for each metadata element 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 and consistency. For example, findings about URLs may be assigned a high, medium or low priority depending on the significance of the finding (***Table 2***). The ARC team has documented the priority matrix for each metadata element and has made the matrices available to the data centers for review (CMR Metadata Best Practices: Landing Page, 2020).

Upon completing all automated and manual assessments for a data center, the findings are compiled in reports and shared directly with the data center (***Figure 2***, step 4). A detailed report is provided for each metadata record assessed, and contains element-by-element recommendations with assigned priority classifications. Detailed reports are meant to be used by discipline-specific data center curators that implement the recommended changes and are also used by the ARC team to track metadata improvements. A summary report is also provided that contains an analysis of common findings and combined metrics for a given set of records. Summary reports are useful to the discipline-specific data center's management staff in estimating the resources needed to address the recommendations and are used by ARC for higher-level reporting. The ARC metadata assessment process concludes when both reports are shared with the data center (***Figure 2***, step 5).

3.3 ASSESSMENT OF UPDATED RECORDS

Upon receiving the ARC assessment reports, each data center formulates a work plan and a corresponding schedule for addressing the report findings. The data center consults with the ARC team to ask any questions about the assessment or to provide feedback on the framework rules. The data center then begins updating metadata in the local database and pushes the improved metadata to the CMR. The ARC team is available to address any questions or concerns that may arise throughout the update process. Once this step is complete, the data center notifies the ARC team that the metadata are ready for reassessment and the ARC team repeats the same quality assessment process for the updated records. Metadata elements that are updated per the ARC recommendations are marked as resolved while elements that have not been updated are reported back to the data center. The data center either continues to work off the ARC team findings or collaborates with the ARC team to come to an agreement regarding findings that the data center does not address. Once all findings are resolved, either through metadata updates or negotiations, the ARC team completes a final metadata quality assessment and compiles change metrics to demonstrate improvements made by the data center.

SECTION 4: PRELIMINARY RESULTS

4.1 MOST PREVALENT HIGH PRIORITY FINDINGS

Preliminary results include assessments for 1,929 datasets in four different metadata standards and from all twelve NASA data centers. Of the records assessed, URLs accounted for the highest number of high priority red findings (***Figure 3***). Typical URL issues include broken URLs and data access URLs that do not conform to NASA security requirements. The ARC team also found that many metadata records did not include links to essential data documentation such as user's guides and ATBDs. Missing documentation is marked as a high priority finding since these documents provide scientific context for users and are therefore important to data usability. More critically, several collection level records did not include data access URLs at all, an omission which prohibits data accessibility and limits the effectiveness of the metadata itself.

The concepts with the second and fourth highest findings, DOI and Collection Progress, were added to the UMM immediately before the ARC team's initial assessment. The Collection Progress element, an indicator of a dataset's production status, is now required by the UMM and was therefore categorized as a high priority finding when missing. Some metadata did not include the Collection Progress element at all while other records provided incorrect Collection Progress values. Most often, the Collection Progress element indicated the data was being actively collected when in fact data collection had been completed. The DOI element, on the other hand, is not required by the UMM but is strongly encouraged to be used for NASA owned data products. While most data centers were requesting DOIs for their datasets and storing those DOIs within the local database, the final step of including the DOI in the CMR metadata

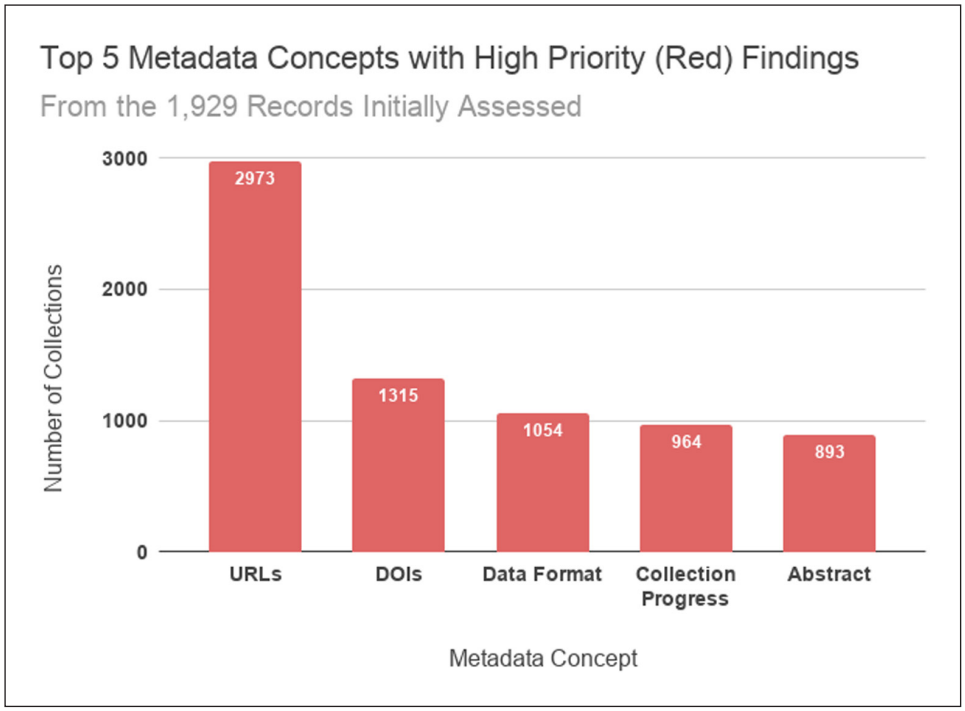

Figure 3 The five collection level metadata concepts that received the most high priority recommendations from the ARC team. Since URLs appear in multiple UMM metadata elements, the number of reported findings shown is more than the number of records reviewed.

had often not yet occurred at the time of initial assessment. These metadata findings suggest that, in some cases, metadata in the CMR is stale and out of sync with the discipline-specific data center. On the other hand, for some data centers, adopting new UMM concepts represents a significant effort due to updates required in the local database.

The remaining high priority findings are related to the Data Format and Abstract elements. These metadata elements are especially important for global catalog users in determining data usability. Complete and easy to understand abstracts are important in determining whether the data is appropriate for a global user's research problem. Similarly, Data Format information helps a user understand whether the data can be quickly and easily incorporated into a research project or workflow. Data Format information was not widely adopted by the data centers because this information was not viewed as critical at the time of initial assessment. Additionally, guidance and consensus on where to include the data format information with the various standards was variable. Once clear guidance was established on data format best practices, several data centers added the Data Format information to their metadata with the end result of this valuable information being used for faceted searches within Earthdata Search.

4.2 METADATA QUALITY IMPROVEMENTS

As of this publication, 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 have been re-evaluated by ARC thus far (***Figure 4***). Between the nine data centers, a total of 19,764 high priority findings, 11,434 medium priority findings and 14,753 low priority findings were reported. Upon reassessment of the nine data centers, the number of findings identified after updates resulted in substantial improvements of 71%, 60%, and 39% in high, medium and low priority findings, respectively. Remaining high, medium and low priority findings have been reported to each data center, and results are expected to continue to improve with subsequent iterations of the metadata assessment process. Iterations will continue until all high priority findings have been addressed.

5. DISCUSSION AND LESSONS LEARNED

5.1 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

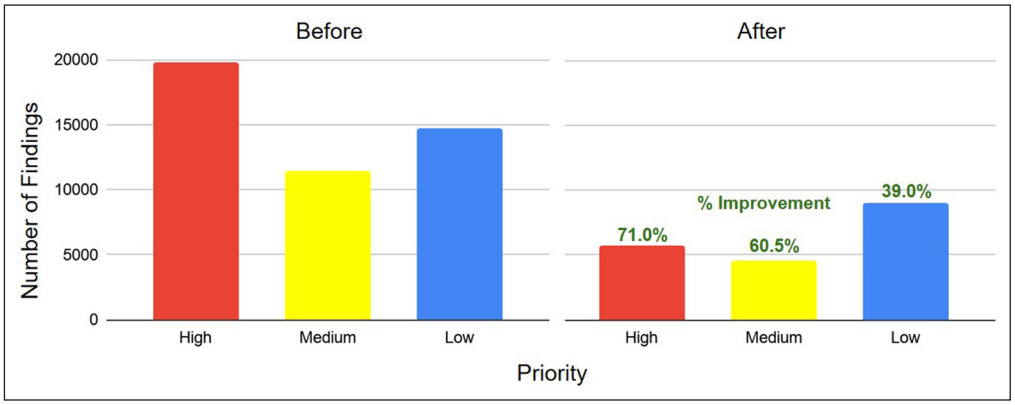

Figure 4 The cumulative number of findings in the high (red), medium (yellow) and low (blue) categories for the nine data centers upon initial assessment (left) and after reassessment (right). The percent improvement in the number of findings is shown above the right three columns.

example, the initial ARC assessments showed that most metadata records use only about 40 percent 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 the content provided can be low, as illustrated in ***Figure 3***.

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) require that all metadata content entered using the tool meet minimum requirements, including compliance with controlled vocabularies. However, metadata can still be incorrect and incomplete even when using a tool. Since most tools only enforce authors to fill in required elements, optional elements are often left blank by users, resulting in incomplete records from an information perspective. In addition, the tool assumes that the metadata author is providing scientifically correct information about the data. Until tools use new techniques to check metadata for scientific correctness, there is always the chance that the resulting metadata could be incorrect. While tools may help with quality, not all data centers opt to use a graphical user interface, preferring to instead work directly with an API or a database. These workflows may leverage schema and keyword validation but also do not check for scientific correctness.

Finally, when bringing together standards into a global catalog, not all standards are equivalent. For example, the CMR supports the integration of five metadata standards into the repository, with a minimum set of required information being met by all of the standards. However, optional content that provides more complete metadata may not be in all five of the standards. In the case of the CMR, some standards are not as well maintained or up-to-date due to long term plans to deprecate the standard. Other standards, such as ISO 19115, support a rich amount of content that can be represented in varied ways which in turn does not easily translate into standards that support fewer concepts. Potential limitations of a given metadata standard, including the degree of maintenance the standard receives or the amount of content the standard supports, should be taken into consideration as part of a discipline-specific data center's stewardship activities.

5.2 METADATA QUALITY AND CONSISTENCY IMPROVES 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). The large scale nature of data production at NASA made metadata standards an essential component in the overall system. 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 SAR and lidar, data from smallsat constellations and airborne and in situ field observations.

For example, airborne and 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. For instance, airborne and field investigations are not made up of continuous observation periods but are instead organized

around field investigations that include one or more deployments (ADMG Airborne and Field Data Inventory Definitions, 2020). This hierarchical investigation structure is not supported in the existing metadata models yet is still important information for data discovery. The lack of flexibility in the model led some of the airborne investigation data centers to attempt to include this information in the metadata using ad hoc approaches. However, each data center implemented a unique approach, leading to inconsistencies across the CMR. Groups like the ARC team and the Airborne Data Management Group (Airborne Data Management Group, 2020) are working with the data centers to build consistent metadata quality approaches until the standards evolve to meet heterogeneous data needs. Longer term, the addition of new metadata elements designed to serve the search and discovery needs of these specific user communities would help ensure consistency across the CMR.

5.3 HIGH LEVEL DATA MANAGEMENT PRINCIPLES AND METADATA QUALITY FRAMEWORKS ARE ONLY EFFECTIVE IF ACTIONABLE RECOMMENDATIONS ARE PROVIDED

Many high level, domain independent data management principles and metadata quality frameworks have been proposed (Bruce & Hillmann 2004; Tani, Candela & Castelli 2013; Wilkinson et al. 2016). However, the effectiveness of these broad principles are limited at best due to the lack of actionable recommendations provided. Principles or recommendations are deliberately open ended to guide the unique implementation choices for each archive. 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 catalogs. When each data center interprets broad principles for individual needs, interoperability and understanding for a broader user community is sacrificed.

A popular example of one of these high level frameworks is the FAIR Data Principles (Wilkinson et al. 2016). The sixteen FAIR Data principles are not a standard, specification or implementation solution, but are meant to guide data centers in implementation choices (Wilkinson et al. 2016). For example, one FAIR principle recommends that 'data are described with rich metadata' (Wilkinson et al. 2016). While the recommendation to provide rich metadata is a good one, it is still too broad and ambiguous for the day-to-day implementation of both authoring and maintaining high quality metadata. In the ARC team's experience, most data centers want to provide high quality metadata that meet both the discipline-specific and global community needs. However, independently determining what rich, high quality metadata looks like resulted in twelve unique interpretations of this principle which are not optimized for interoperability across the global system. The ARC team's definitive and independent recommendations, along with the priority classification of those recommendations, have provided the concrete guidance needed to make effective metadata quality changes easier for the data centers.

5.4 HIGH QUALITY METADATA DOES NOT GUARANTEE CONSISTENT DATA ACCESSIBILITY

High quality metadata increases the discoverability of data and also raises the likelihood that data will be accessible and usable. However, high quality metadata does not guarantee consistent data accessibility and usability. For example, data access methods across NASA's twelve data centers are variable, making a consistent data access experience impossible. Across the twelve NASA data centers, there are more than 61 data tools available for searching and ordering, subsetting, filtering, reprojection, geolocation and data visualization (Liu et al. 2020). Some of the data centers rely entirely on these ordering tools for data access while other data centers provide direct file access. For those data centers that provide direct file access, the file access structure varies, making it challenging for a user to easily find the correct data. Similarly, the ARC team found that data services such as the Open-source Project for a Network Data Access Protocol (OPeNDAP) were configured in different ways across the data centers, again making a consistent data access experience difficult. 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 makes a consistent data access experience unattainable at this time.

5.5 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 as 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 require that the issues be communicated to the discipline-specific data center and resolved at the source in the local database. Updated records must then be pushed to the aggregated catalog; otherwise no modifications are made to the metadata records within the aggregated catalog, and global metadata quality is not improved. While this operational philosophy is a good best practice for maintaining metadata quality across multiple databases, we suggest that the aggregator role should be expanded to enable metadata quality in collaboration with the discipline-specific data centers. The aggregator should be empowered to make changes to metadata in the global environment but, in the interest of trustworthiness and transparency, should also communicate those changes back to the discipline-specific data centers in order for metadata quality to be maintained across the enterprise.

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 (Hillmann 2008). These changes include resolving typographical errors, removing deprecated elements (Hillmann 2008) and detecting duplicate records (Stvilia, Gasser & Twidal 2004). Transformation changes are designed to only correct existing metadata, and should not replace or modify existing metadata that is already correct nor should it affect the semantics of the record. Transformation changes are easier to adopt for aggregators since these changes are essentially enhancing existing metadata and not fundamentally changing the content. Augmentation changes, on the other hand, leverage the information provided in the metadata to add new information or value to the aggregated record (Hillmann 2008). These additions include, but are not limited to, detecting and adding the data format to the metadata, using machine learning techniques to add scientific keywords or topics to the metadata, gathering and leveraging collection statistics (Stvilia, Gasser & Twidal 2004). Augmentation changes may be adopted on an organization by organization basis and may be more appropriate for more organizationally aligned aggregators like the CMR. The 'orchestra' of automated and manual augmentation services (Tani, Candela & Castelli 2013) increases the exposure of data in an aggregated environment and the likelihood that data will be reused. Finally, to ensure that metadata quality is maintained across the enterprise, the aggregator may consider using automated techniques to communicate quality changes back to the discipline-specific centers so that metadata may be updated.

6. CONCLUSIONS

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. In this paper, the ARC team has demonstrated that metadata quality in an aggregated catalog 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 experiences, 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. Once the metadata has been evaluated, close collaboration with each of the NASA data centers is an essential aspect to successfully implementing the ARC metadata quality recommendations. This close collaboration requires clear, consistent and open communication with collaborators and stakeholders throughout the process. Beyond collaborating with the NASA data centers, collaboration with the broader data system team was also necessary for providing feedback about needed systematic changes, such as updates to the metadata models, that emerged as a result of the quality assessment process.

Providing an independent perspective to the metadata quality assessment process is also beneficial. ARC's situation as neutral metadata quality reviewers has benefited the overall

process by providing recommendations that assess metadata from an enterprise, data system perspective instead of an archive-by-archive approach. While most data centers excel at considering the needs of discipline-specific users, the needs of global users may not always be prioritized due to the ambiguity surrounding those needs. The ARC team process reduces the ambiguity between discipline-specific and global user needs for the data centers by providing easy to understand and actionable recommendations. These recommendations also provide a common goal that all of the data centers may together achieve. Several data centers expressed a desire to improve metadata quality for global users but were uncertain where to begin. The ARC recommendations were welcomed by these data centers as a criterion for coordinated and strategic metadata quality improvements. The ARC metadata assessment process shows that data centers want to improve metadata quality but prefer to wait for guidance so as to maintain consistency with the broader community. On the other hand, ARC's process has shown that, for some data centers, an organizational mindset shift was needed from focusing solely on discipline-specific users to a more inclusive, global-oriented mindset.

Data systems and archives who wish to implement a similar metadata quality assessment process should consider the following recommendations derived from the ARC process. First, establish metadata quality priorities, such as data discovery, accessibility or provenance, as early as possible in the process. These priorities guide all of the detailed recommendations so agreeing on the goal of the quality assessment process before moving to detailed recommendations is essential. Second, be as transparent as possible with all stakeholders as early as possible in the process. This emphasis on transparency includes providing open access to detailed metadata quality assessment checks and ensuring collaboration and feedback from stakeholders is ongoing. Third, be as flexible and adaptable as possible when creating and maintaining quality assessment checks. Establishing and refining metadata quality checks is an iterative process that will evolve as stakeholders provide more feedback. In addition, metadata models evolve to address changing needs which, in turn, drives changes to the assessment checks. Fourth, recognize that time and resources are limited at both the individual data center level and more broadly at the enterprise level. Providing prioritized recommendations helps all stakeholders determine which findings to address first and to also draft work plans to support metadata quality activities. Last, the ARC team has made a number of metadata quality resources available for reuse by the community. These resources include documentation of the ARC team's metadata quality checks for each element, an open-sourced metadata quality review tool (CMR Metadata Review Tool, 2021) and scripts of automated checks based on the ARC framework (pyQuARC, 2021). In addition to basic validation checks, the scripts flag opportunities to improve or add contextual metadata information and also ensure that information common to both collection and the granule-level metadata are consistent and compatible.

We recognize that the ARC methodology is not a long-term sustainable solution for maintaining metadata quality. While the process is systematic and thorough, it still relies heavily on manual efforts which are time and resource intensive. Instead, we envision integrating new techniques, such as artificial intelligence (AI) and machine learning (ML), into all stages of the metadata curation lifecycle. For example, the team is currently prototyping a tool which uses machine learning techniques to compare a dataset abstract to a corpus of scientific literature in order to recommend consistent scientific keywords for metadata. Techniques such as these will not only improve the initial creation of metadata but will also assist in monitoring metadata quality within the system over time. As science continues to be more and more data driven, we recognize that data discovery needs, along with the corresponding metadata models, will change. These changes will require data systems to prototype emerging new technologies, such as graph databases and cloud-based data platforms, in order to enable these new discovery paradigms. Most likely, metadata quality needs will change as these new paradigms are adopted, again requiring stakeholders to be flexible and adaptable. Lastly, as more data systems work to support open science, metadata needs will inevitably expand to support other first class research objects such as software and documentation. Combining new AI/ML techniques, new technological solutions and lessons learned from metadata quality assessment projects like ARC should make the metadata expansion required for open science easier and more sustainable.

ACKNOWLEDGEMENTS

We would like to thank Deborah Smith and Derek Koehl for assistance in improving the manuscript. We would also like to thank all our colleagues at the DAACs, the ESDIS Project and the EED2 metadata quality team, for their continued collaboration and support.

FUNDING INFORMATION

This directed work was completed in support of NASA's Earth Science Data Systems program.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Kaylin Bugbee *orcid.org/0000-0001-6733-5698*
NASA, US

Jeanné le Roux *orcid.org/0000-0002-8274-987X*
University of Alabama in Huntsville, US

Adam Sisco *orcid.org/0000-0002-5974-3402*
University of Alabama in Huntsville, US

Aaron Kaulfus *orcid.org/0000-0002-8319-1126*
NASA, US

Patrick Staton *orcid.org/0000-0002-9415-2413*
University of Alabama in Huntsville, US

Camille Woods *orcid.org/0000-0002-7586-9021*
University of Alabama in Huntsville, US

Valerie Dixon *orcid.org/0000-0002-5125-2270*
NASA, US

Christopher Lynnes *orcid.org/0000-0001-6744-3349*
NASA, US

Rahul Ramachandran *orcid.org/0000-0002-0647-1941*
NASA, US

REFERENCES

ADMG Airborne and Field Data Inventory Definitions, 5 August 2020. Available at *https://earthdata.nasa.gov/esds/impact/admg/admg-definitions* [Last accessed 21 December 2020].

Alaska Satellite Facility (ASF). Available at *https://asf.alaska.edu/* [Last accessed 21 December 2020].

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*.

Baynes, K and **Mitchell, A** 2017. The Common Metadata Repository: The Foundation of NASA's Earth Observation Data, 1 March 2017. Available at *https://earthdata.nasa.gov/learn/articles/the-common-metadata-repository* [Last accessed 21 December 2020].

Borgman, CL. 1986. Why are online catalogs hard to use? Lessons learned from information-retrieval studies. *Journal of the American Society for Information Science*, 37(6): 387–400. DOI: *https://doi.org/10.1002/(SICI)1097-4571(198611)37:6<387::AID-ASI3>3.0.CO;2-8*

Brown, ME, et al. 2013. Policy for robust space-based earth science, technology and applications. *Space Policy*, 29(1): 76–82. DOI: *https://doi.org/10.1016/j.spacepol.2012.11.007*

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*.

CDDIS. Available at *https://cddis.nasa.gov/* [Last accessed 21 December 2020].

CMR Metadata Best Practices: Landing Page. Available at *https://wiki.earthdata.nasa.gov/display/CMR/CMR+Metadata+Best+Practices%3A+Landing+Page* [Last accessed 21 December 2020].

CMR Metadata Review Tool. Available at *https://github.com/nasa/cmr-metadata-review* [Last accessed 14 April 2021].

Data.gov. Available at *https://www.data.gov/* [Last accessed 21 December 2020].

Earthdata Search. Available at *https://search.earthdata.nasa.gov/search* [Last accessed 21 December 2020].

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 Cyber infrastructures."* Available at *https://deepblue.lib.umich.edu/bitstream/handle/2027.42/49353/UnderstandingInfrastructure2007.pdf?sequence=3&isAllowed=y*.

Edwards, PN, et al. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5): 667–690. DOI: *https://doi.org/10.1177/0306312711413314*

***FedEO.** Available at *https://eoportal.org/web/eoportal/fedeo* [Last accessed 21 December 2020].

Hillmann, DI. 2008. Metadata Quality: From Evaluation to Augmentation. *Cataloging & Classification Quarterly*, 46(1): 65–80. DOI: *https://doi.org/10.1080/01639370802183008*

Khalsa, SJS, et al. 2011. Towards Unifying NASA Earth Science Enterprise-Wide Metadata Around International Standards: Study Results and Recommendations. *28th International Symposium on Remote Sensing of Environment*. Available at *https://www.isprs.org/proceedings/2011/ISRSE-34/211104015Final00852.pdf*.

Liu, Z, et al. 2020. Creating Data Tool Kits That Everyone Can Use. *Eos*, 101. DOI: *https://doi.org/10.1029/2020EO143953*

NASA. 2017. *EOSDIS Handbook Version 1.3*. Available at *https://cdn.earthdata.nasa.gov/conduit/upload/6321/EOSDIS_Handbook_1.3.pdf*.

Overpeck, JT, et al. 2011. Climate data challenges in the 21st century. *Science*, 331(6018): 700-702. DOI: *https://doi.org/10.1126/science.1197869*

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*

pyQuARC: Open Source Library for Earth Observation Metadata Quality Assessment, 9 April 2021. Available at *https://github.com/NASA-IMPACT/pyQuARC#readme*.

Ramachandran, R, et al. 2016. Climate data initiative: A geocuration effort to support climate resilience. *Computers and Geosciences*, 88: 22–29. DOI: *https://doi.org/10.1016/j.cageo.2015.12.002*

Stvilia, B, et al. 2007. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12): 1720–1733. DOI: *https://doi.org/10.1002/asi.20652*

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: *http://doi:10.4018/978-1-59904-420-0.ch008*

Tani, A, Candela, L and **Castelli, D.** 2013. Dealing with metadata quality: The legacy of digital library efforts. *Information Processing and Management*, 49(6): 1194–1205. DOI: *https://doi.org/10.1016/j.ipm.2013.05.003*

The Airborne Data Management Group (ADMG). 25 August 2020. Available at *https://earthdata.nasa.gov/esds/impact/admg* [Last accessed 21 December 2020].

The Cambridge Crystallographic Data Centre (CCDC). Available at *https://www.ccdc.cam.ac.uk/* [Last accessed 21 December 2020].

White House Office of Science and Technology Policy (OSTP). 2016. *Common Framework for Earth-Observation Data*. Available at: *https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/common_framework_for_earth_observation_data.pdf*.

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*

World Data Center for Climate (WDCC). Available at *https://www.dkrz.de/up/systems/wdcc* [Last accessed 21 December 2020].

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*

TO CITE THIS ARTICLE:
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: 17, pp. 1–15. DOI: *https://doi.org/10.5334/dsj-2021-017*

Submitted: 29 December 2020
Accepted: 11 April 2021
Published: 28 April 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.