pyQuARC: Preparing for Full Release

IMPACT
Interagency Implementation and Advanced Concepts Team

The Importance of Metadata Quality

Metadata provides the structured contextual information needed by applications to effectively connect users to data. As a result, high quality metadata are needed to maximize data discoverability and accessibility. IMPACT's Analysis and Review of the CMR (ARC) team assesses the completeness, correctness, and consistency of NASA's Earth Observation metadata records in the Common Metadata Repository (CMR) to ensure that the data are accessible, usable, and discoverable. To streamline the review process, the team developed a suite of python scripts called **pyQuARC** that uses ARC's existing metadata quality framework to provide prioritized recommendations for metadata improvement.

Automating Metadata Quality Assessments with pyQuARC

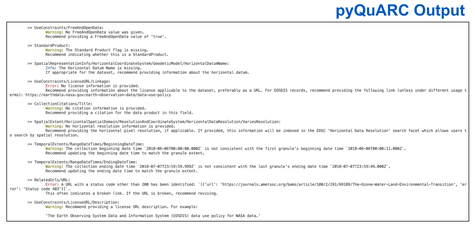
pyQuARC is a customizable, open-source library that completes basic validation checks, pinpoints inconsistencies between dataset-level (collection) and file-level (granule) metadata, and identifies opportunities for more descriptive and robust information. The metadata formats supported by pyQuARC include those supported by the CMR: **DIF10** (collection only), **ECHO10** (collection and granule), **UMM-C**, and **UMM-G**. Findings are categorized by priority level based on the following criteria:

ARC Priority Matrix

High [Error]	• Inaccurate, incomplete, or missing content • Broken URLs and invalid collection-granule relationships
Medium [Warning]	• Revisions of existing content • Addition of new information
Low [Info]	• Minor consistency issues

When pyQuARC is executed on a metadata record, the user receives the findings as output with the prioritized color code preserved (shown below).

pyQuARC Output



Authors: Essence Raphael[1], Slesa Adhikari[1], Ashish Acharya[1], Rajesh Pandey[1], Iksha Gurung[1], Jeanné le Roux[1], Jenny Wood[1], Sheyenne Kirkland[1], Aaron Kaulfus[2], Kaylin Bugbee[2], Rahul Ramachandran[2]

[1]UAH [2]NASA MSFC

Contact: essence.raphael@uah.edu

This work is supported by NASA Grant 80MSFC22M004.

After a year of improvements, pyQuARC automatically flagged ~74% of metadata findings from a sample of manually reviewed records

Access pyQuARC here!

The pyQuARC library is currently in beta phase. In preparation for its full release, the ARC team has completed performance testing to examine how well pyQuARC's automated reviews capture the recommendations of ARC team members. Initial testing showed that pyQuARC automatically flagged 58% of metadata findings from a sample of manually reviewed records. Based on these results, the ARC team implemented a number of improvements to pyQuARC.

Checks Added	Additional Modifications
• Science Keyword Presence Check • DOI Authority Presence Check • Free and Open Data Check • Horizontal Resolution Presence Check • Data Format Presence Check • Standard Product Check • License URL Description Check • Granule Campaign Name Presence Check	• Updated schema files • Added fields to the Spatial Extent Fulfillment Check • Revised GCMD Long Name Presence Checks • Revised the Validate Beginning Datetime Against Granules Check • Modified output messages to better communicate the recommended changes/actions

Testing results for a sample metadata record

Concept ID: ID for the metadata record

Schema: Name of metadata schema

Field/Check	ARC	pyQuARC	Matches	Inconsistencies	Notes
Free and Open Data	✔	✔	1	0	
Standard Product	-	✔	0	1	
Horizontal Datum Name	✔	✔	1	0	
License URL Linkage	✔	✔	1	0	
Collection Citation	✔	✔	1	0	
Horizontal Data Resolutions	✔	✔	1	0	
Beginning Date Time	-	✔	0	1	
Ending Date Time	✔	✔	1	0	
Processing Level Description	✔	✔	1	0	
Related URLs Description	✔	✔	1	0	
Platform Long Name	✔	✔	1	0	
License URL Description	✔	✔	1	0	
Abstract - grammar/acronyms	✔	-	0	1	No pyQuARC functionality
Location Keywords Category	✔	-	0	1	No pyQuARC functionality
		Totals:	10	4	

Performance Results:

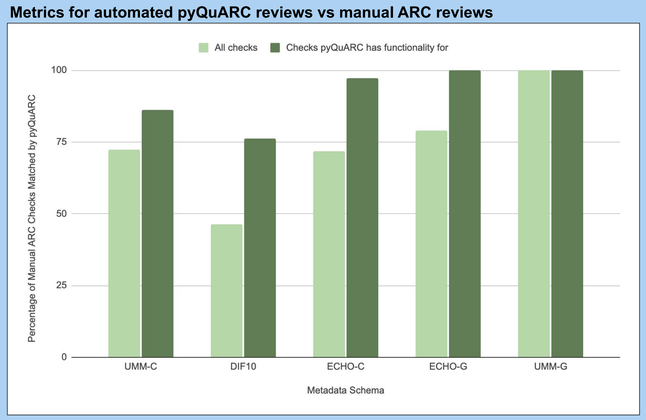
71% match with ARC findings

83% match with ARC findings with pyQuARC functionality

To assess the effectiveness of the most recent set of changes, the ARC team compared findings from a manual metadata review and an automated pyQuARC review. This was done for a number of metadata records from each supported metadata schema and aimed to answer the question, "*For a given metadata field, did pyQuARC flag the same issues as the ARC reviewer?"* The results showed that pyQuARC flagged **~74%** of metadata findings from manual reviews, illustrating the improvements made since initial testing.

Some of the discrepancies between pyQuARC's automated reviews and those done manually exist due to the many additional recommendations that a manual ARC reviewer can identify (e.g. grammar/misspellings, lack of descriptive abstracts, etc.) that pyQuARC does not have the ability to detect. When considering only the recommendations pyQuARC can identify, it flagged **~92%** of findings from manual reviews. These results are broken down by schema in the bar graph below.

Metrics for automated pyQuARC reviews vs manual ARC reviews



There are known pyQuARC bugs and nuances in the metadata that also contribute to the results for each schema. Notable issues include:

UMM-C	• Issue reading the 'Standard Product' field • Flagging links for sites that require authentication • Issue handling differences in temporal extent precision between the collection- and granule-level metadata
DIF10	• Issue displaying 'Use Constraints' findings • Flagging links for sites that require authentication • Issue handling extra 'Data Resolution' metadata tag
ECHO-C	• Issue handling differences in temporal extent precision between the collection- and granule-level metadata

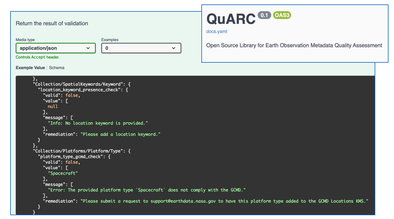
The high percentage of manual findings pyQuARC is able to capture (for functional checks) demonstrates that **pyQuARC development is nearing completion**. The ARC team is actively working to resolve the remaining bugs and nuances, and additional features such as sub-field recommendations (e.g. OPeNDAP subfields in granule-level metadata) may be added as well.

Additional Use Cases

pyQuARC not only assists the ARC team with more effectively reviewing the quality of metadata records published in the CMR but can also be used by metadata providers for maintenance, providing on-demand, automatic quality assessments of their metadata. For example, one such metadata provider is currently using pyQuARC to obtain updated assessments on metadata records that ARC has reviewed in the past as well as assessments for new records. Because pyQuARC is customizable, it can also be tailored to check SpatioTemporal Asset Catalogs (STAC) and International Organization for Standardization (ISO) formatted Earth observation metadata. The pyQuARC GitHub repository provides instruction for customizing pyQuARC.

QuARC: pyQuARC as a Service

pyQuARC is also available as a cloud-based service called **QuARC** where metadata records can be reviewed through APIs, facilitating easier adoption and integration into workflows.



pyQuARC's Full Release

pyQuARC's transition out of the beta phase is tentatively scheduled for **March 2024**. Although pyQuARC will not be able to provide the same level of detail as an ARC reviewer, this process of testing, revising, and enhancing the tool will ensure that the assessments pyQuARC provides are accurate and reliable with as much automated functionality as possible. As the ARC project moves towards closeout, pyQuARC can potentially be leveraged as a viable replacement for manual quality assessments and may eventually be maintained by the broader EOSDIS community.

Resources

- pyQuARC on GitHub: https://github.com/NASA-IMPACT/pyQuARC
- ARC Metadata Quality Framework publication (Bugbee et al. 2021): https://doi.org/10.5334/dsj-2021-017
- NASA Earth science data can be found and downloaded at NASA Earthdata Search: https://search.earthdata.nasa.gov/search
- Your Gateway to NASA Earth Observation Data, learn more at: https://www.earthdata.nasa.gov/

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

NASA