QuARC: Development of a Service to Enable FAIR-er Metadata

IMPACT
Interagency Implementation and Advanced Concepts Team

The ARC Project

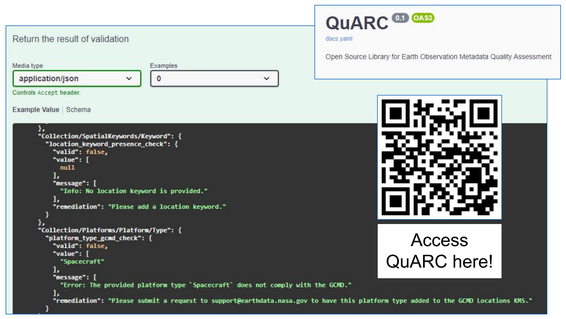
The ARC Team located at NASA's Marshall Space Flight Center conducts quality assessments of metadata records that catalog NASA's collection of over 9,000 Earth observation data products, stored in a centralized database called the Common Metadata Repository (CMR). The ARC Team has developed a metadata quality assessment framework to evaluate metadata completeness, correctness, and consistency with the goal of making NASA's data products more discoverable, accessible, and usable.
ARC = Analysis and Review of the CMR

Automating Metadata Quality Assessments

To streamline this process, ARC has developed a suite of python scripts called **pyQuARC**: an open-source python library for Earth Observation Metadata Quality Assessment. pyQuARC is designed around ARC's metadata quality assessment framework to make basic validation checks, pinpoint inconsistencies between dataset-level (i.e. collection) and file-level (i.e. granule) metadata, and identify opportunities for more descriptive and robust information. Since pyQuARC is customizable, users can make modifications as needed.

QuARC: pyQuARC as a Service

- pyQuARC has now been deployed on an Amazon Web Services cloud environment (QuARC) with an API, allowing for easier adoption of the automated checks.



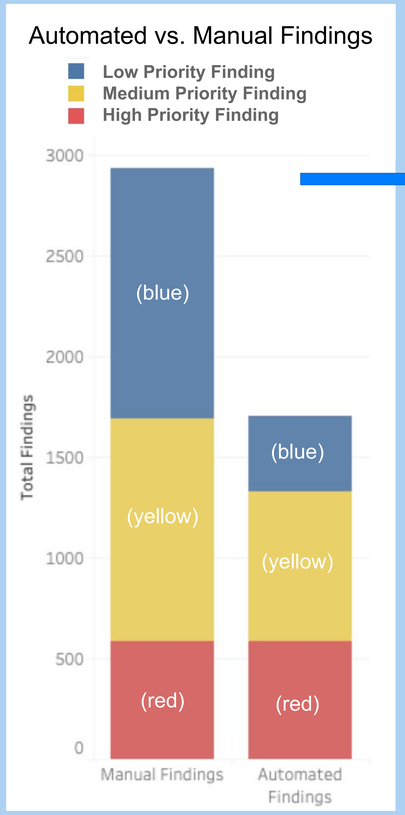
Authors: Jeanné le Roux[1], Slesa Adhikari[1], Ashish Acharya[1], Rajesh Pandey[1], Iksha Gurung[1], Samuel Ayers[1], Jenny Wood[1], Shelby Bagwell[1], Essence Raphael[1], Stephen McNeal[1], Aaron Kaulfus[1], Danielle Groenen[1], Kaylin Bugbee[2], Rahul Ramachandran[2]

[1]UAH [2]NASA MSFC

Contact: jeanne.leroux@uah.edu

This work is supported by NASA Grant 80MSFC22M004.

In an initial test, pyQuARC automatically flagged 58% of metadata findings from a sample of manually reviewed records



ARC ran a set of 180 CMR metadata records from a diverse set of campaigns through pyQuARC. We then compared pyQuARC's results to ARC's manual findings for the same set of records. The results of this initial test are shown in the figure on the left.

Resulting metrics show that pyQuARC caught 58% of ARC's recommendations. Breaking that down further, pyQuARC identified approximately 100% of ARC's red recommendations, 67% of ARC's yellow recommendations, and 30% of ARC's blue recommendations.

Digging deeper

- Next, the team analyzed the manual recommendations from each record against the pyQuARC output. The table and description below summarizes our findings from this detailed analysis:

Working Correctly	Recommendations Missed by pyQuARC	Recommendations Missed by ARC
• GCMD compliance and other controlled vocabulary checks • Field Presence checks • Broken link checks • Duplication checks	• Additions to make the metadata more robust (e.g. adding science keywords) • Removing redundant or unnecessary information • Spelling/grammatical errors	• Providing citation information • Providing more detailed processing level descriptions • Providing a spatial coverage type • Removing duplicate URL descriptions • Providing a "GET DATA" link

- **Recommendations missed by pyQuARC:** Most findings that pyQuARC missed are difficult to automate, for example, identifying science keywords that may be missing or grammatical errors. This is an area where the team can explore possibilities for further automation.
- **Recommendations missed by ARC:** This exercise highlighted several examples where a manual reviewer missed a finding. This illustrates the benefit of pyQuARC in assisting manual reviewers who have to keep track of a long checklist of quality criteria.
- **Nuances in manual checks:** This exercise demonstrated a few cases where pyQuARC and manual reviewers identified the same finding, but assigned it a different priority color. For example, pyQuARC consistently flags temporal inconsistencies as yellow (medium priority), but a manual reviewer can further distinguish why a temporal inconsistency is present and categorize it accordingly (e.g. high or low priority).
- **Checks that could be added:** Since development began, metadata models have been updated resulting in some additional checks that need to be added to pyQuARC. For example, "Free and Open Data" is a new field currently not checked by pyQuARC, but a check can be added to evaluate whether the flag is present and is a boolean value.
- **Bugs:** Some known bugs exist within the pyQuARC infrastructure. The ARC team is actively working to resolve these to finalize development of the tool. For example, the "temporal granule consistency check" produces an error that needs additional investigation, and this did contribute to a lower % match in the results.

Additional Use Cases

In addition to helping the ARC team more quickly and effectively complete metadata assessments, QuARC is also being considered by other teams. For example, NASA's Commercial Smallsat Data Acquisition (CSDA) program actively utilizes QuARC to monitor the metadata catalog of commercial data holdings. Some NASA Distributed Active Archive Centers (DAACs) have also expressed interest in utilizing QuARC as an additional validation step prior to publishing metadata to the CMR. While not currently supported, QuARC can be expanded to support checking of other popular metadata standards and formats such as SpatioTemporal Asset Catalogs (STAC). Please refer to the pyQuARC GitHub repository (linked under 'Resources') for additional information on how to customize and add metadata checks.

FAIR-er Metadata in the Future

The ARC project is dedicated to improving metadata, contributing to thousands of improvements and corrections in NASA metadata (Bugbee et al. 2021), and promoting the FAIR data principles (Finable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016). The long term vision for pyQuARC is community adoption and ownership. Community contributions to pyQuARC will ensure FAIR-ness as metadata and community standards evolve. QuARC provides the means for more widespread adoption and integration into tools and workflows.

Resources

- pyQuARC on GitHub: https://github.com/NASA-IMPACT/pyQuARC
- ARC Metadata Quality Framework publication (Bugbee et al. 2021): http://doi.org/10.5334/dsj-2021-017
- NASA CSDA: https://www.earthdata.nasa.gov/esds/csda
- FAIR paper (Wilkinson et al. 2016): https://doi.org/10.1038/sdata.2016.18
- 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