Collaborative Metadata Curation in Support of NASA Earth Science Data Stewardship

2018 ESDSWG Meeting
Annapolis, MD
April 19, 2018

Adam W. Sisco[1], Kaylin Bugbee[1], Jeanne le Roux[1], Patrick Staton[1], Brian Freitag[1], and Valerie Dixon[2]

(1) University of Alabama in Huntsville, (2) NASA Goddard Space Flight Center, ESDIS

THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

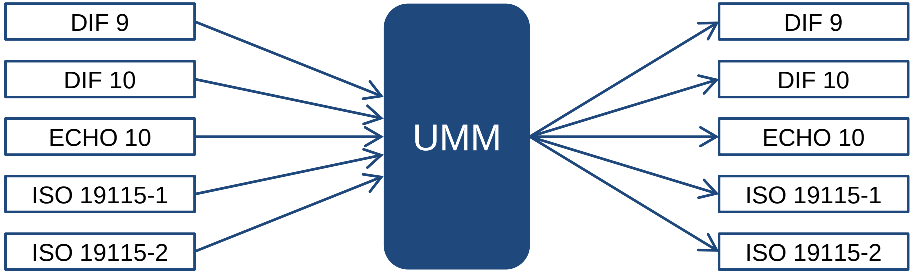
EOSDIS and CMR

- Growing collection of NASA Earth science data is archived and distributed by EOSDIS's 12 Distributed Active Archive Centers (DAACs)

Collections	Granules
6,964	380M

- Each collection and granule is described by a metadata record housed in the Common Metadata Repository (CMR)

- Multiple metadata standards are in use, and core elements of each are mapped to and from a common model – the Unified Metadata Model (UMM)

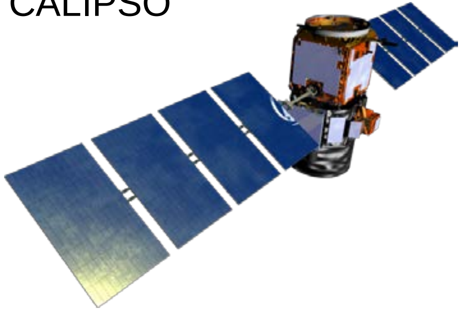


Earthdata Search

- The Earthdata Search Client uses metadata in the CMR to **present users with the information they are looking for and hand users off to more specific applications**
  - Are users finding the information they are looking for? If not, why?
  - Are users being handing off to more specific applications? If not, why?
- Poor quality metadata is often the answer
- The CMR functions best when the metadata it houses is complete, consistent, and accurate
- Let’s examine real examples of “less than ideal” metadata and the consider the consequences

Discovery

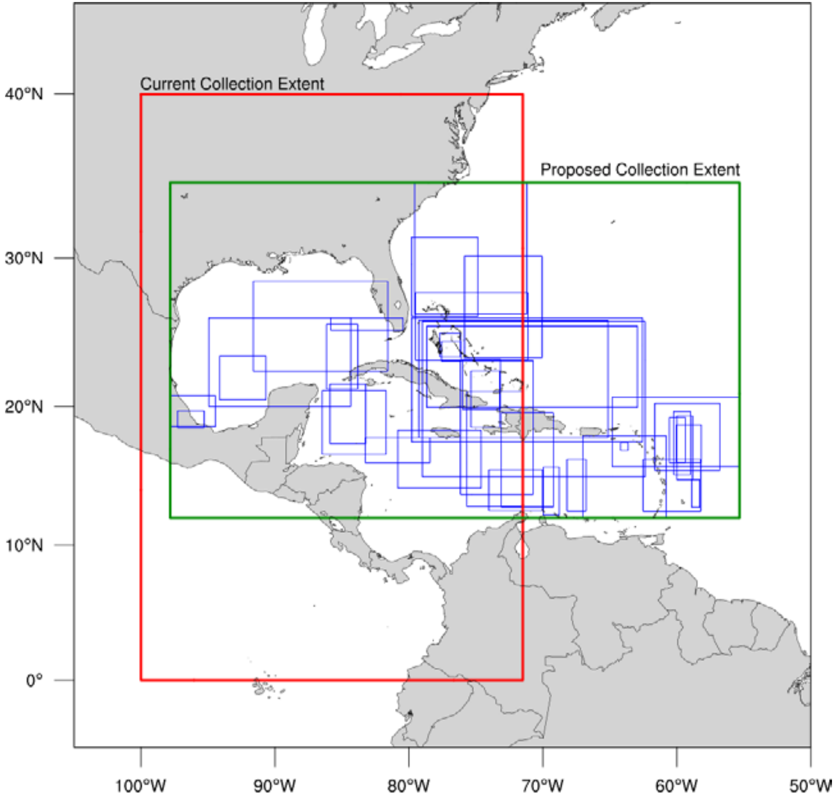
CALIPSO



- Wide Field Camera (WFC) → 171K granules
- Imaging Infrared Radiometer (IIR) → 450K granules
- Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) → 1 granule

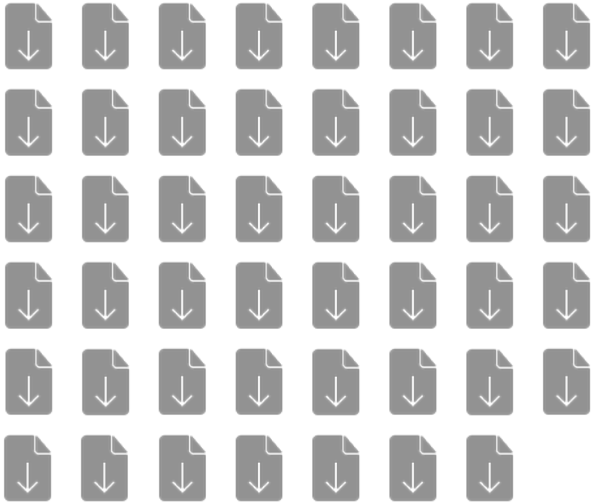
LIDAR → 2M granules

GRIP Field Experiment



Accessibility

- Can I access the data via direct download?
- Served correct data?
- Served all data requested?



47 granules

19 are not published to CMR

Usability

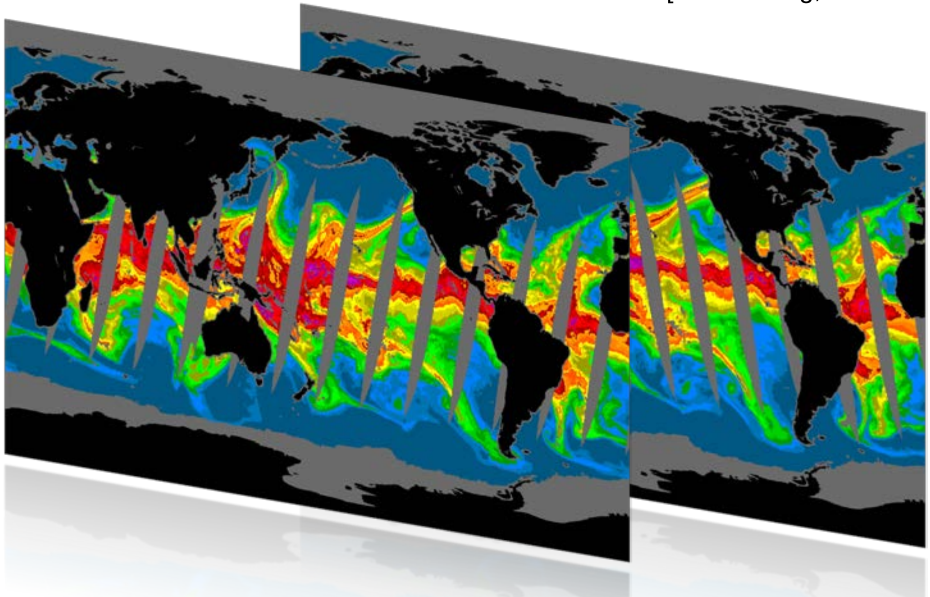
- Are users presented with the option to be handed off to online documentation?

- Data set landing pages
- User’s guides
- README files
- Algorithm Theoretical Basis Documents
- FAQ pages
- Data recipes, how-to guides, tutorials
- Related journal publications
- Quality assessments

- Verify accuracy of metadata and documentation, especially for highly visible collections

User’s guide and netCDF global attributes: [Ascending, Descending]

File is structured: [Descending, Ascending]



What is metadata curation?

Traditional curation

Information Age web content curation

Digital curation

"Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle."

Digital Curation Center, Edinburgh, Scotland

Analysis and Review of CMR (ARC) Team

- Team of several current and former users of NASA Earth science data for research applications

- Science backgrounds in:
  - Earth science
  - Atmospheric science
  - Space science
  - Remote sensing

- Previous curation experience from the Climate Data Initiative (CDI)
  - Review of 850 metadata records for quality and accessibility

ARC's Approach to Digital Curation

Automated Compliance Review

- Ensures elements required by the UMM are populated
- Verifies compliance with controlled vocabularies and native schema enumerations
- Reports state of URLs
- Checks that DOIs are present and resolvable
- Flags lack of data format information
- Identifies invalid collection-granule relationships
  - Temporal coverage
  - Spatial coverage

ARC’s Approach to Digital Curation

Manual Content Review

- Accuracy
  - Transposition of information
  - Invalid platforms and instruments
- Addition of information supported by the model
  - Geodetic model
  - Spatial resolution
  - Related publications
  - Science keywords
  - Data format
  - Citation information
- Consistency, comprehensibility, keyword relevancy
- Access to data and documentation

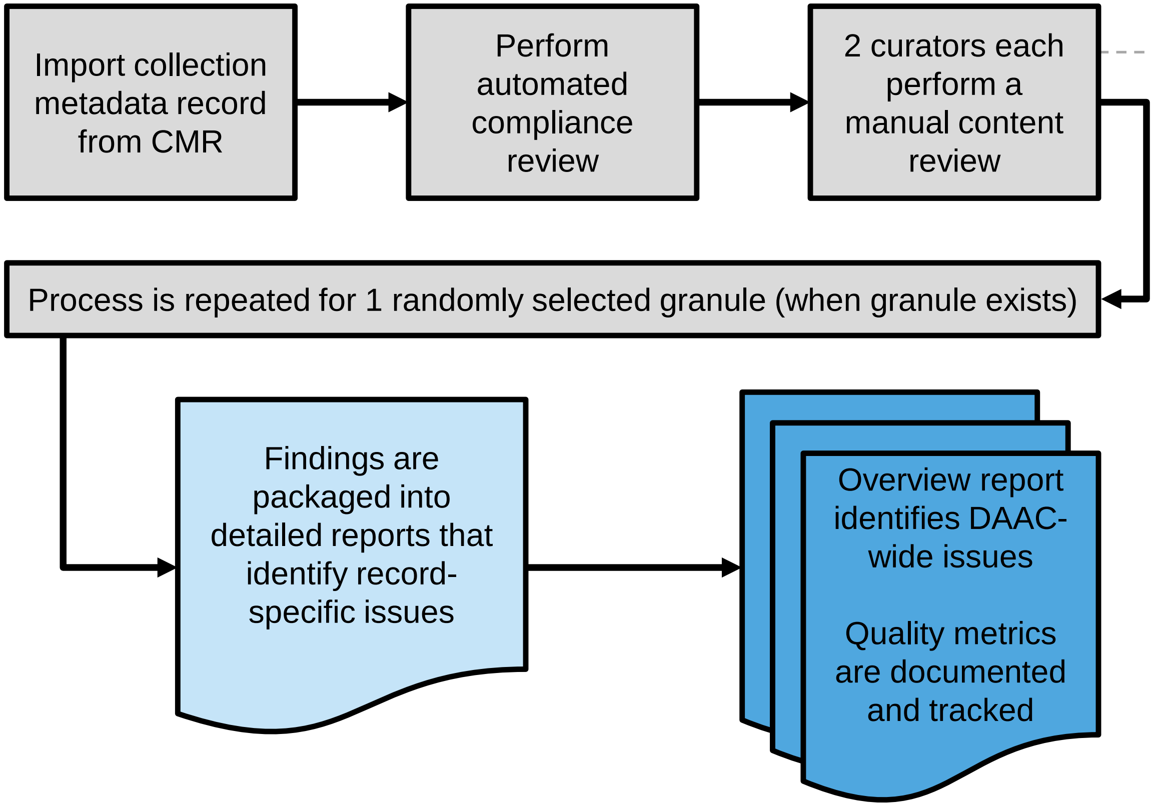
Did I get lost along the way? Could the number of clicks it takes to get to the data and pertinent information be reduced?

DMP-1/2

What else might I need to get started with these data (especially binary)?

DMP-4

ARC Curation Process



Import collection metadata record from CMR

Perform automated compliance review

2 curators each perform a manual content review

Process is repeated for 1 randomly selected granule (when granule exists)

Findings are packaged into detailed reports that identify record-specific issues

Overview report identifies DAAC-wide issues

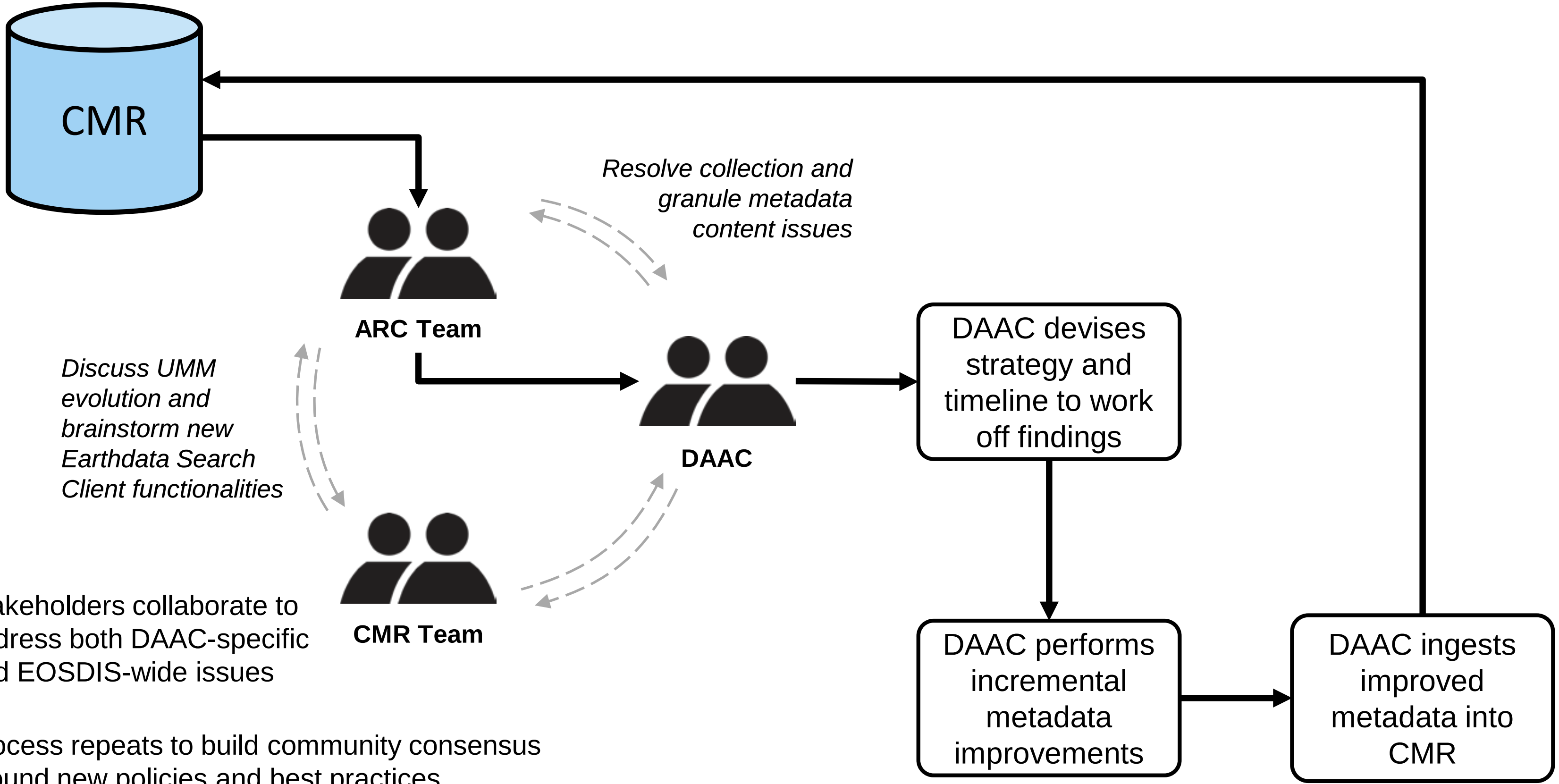
Quality metrics are documented and tracked

High	• Inaccurate, incomplete, or missing content • Broken URLs and invalid collection-granule relationships
Med	• Revisions of existing content • Addition of new information
Low	• Minor consistency issues

Priority classification scheme

1. Assists DAAC in formulating a strategic plan to address findings
2. Used to track resolution of issues

ARC Curation Process



CMR
ARC Team
DAAC
CMR Team
DAAC devises strategy and timeline to work off findings
DAAC performs incremental metadata improvements
DAAC ingests improved metadata into CMR

Resolve collection and granule metadata content issues

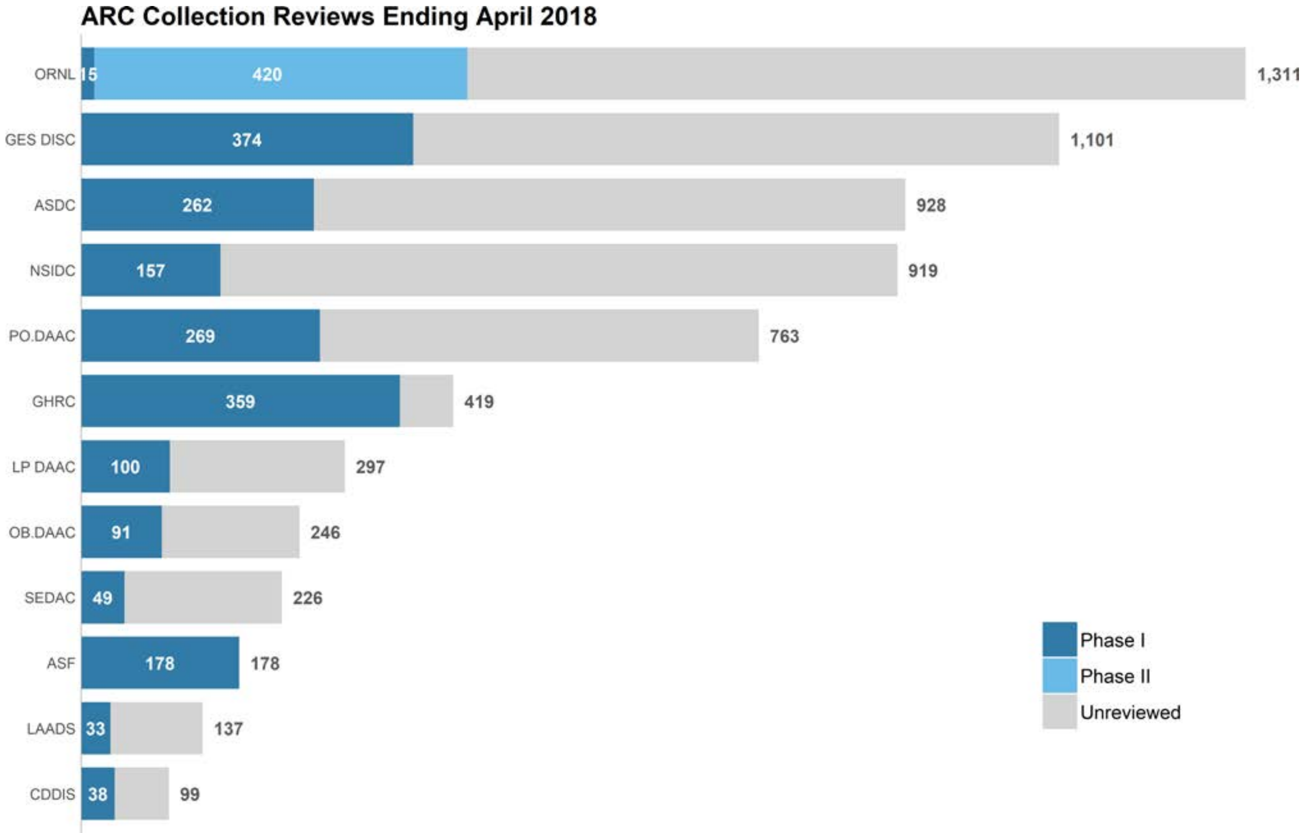
Discuss UMM evolution and brainstorm new Earthdata Search Client functionalities

Stakeholders collaborate to address both DAAC-specific and EOSDIS-wide issues

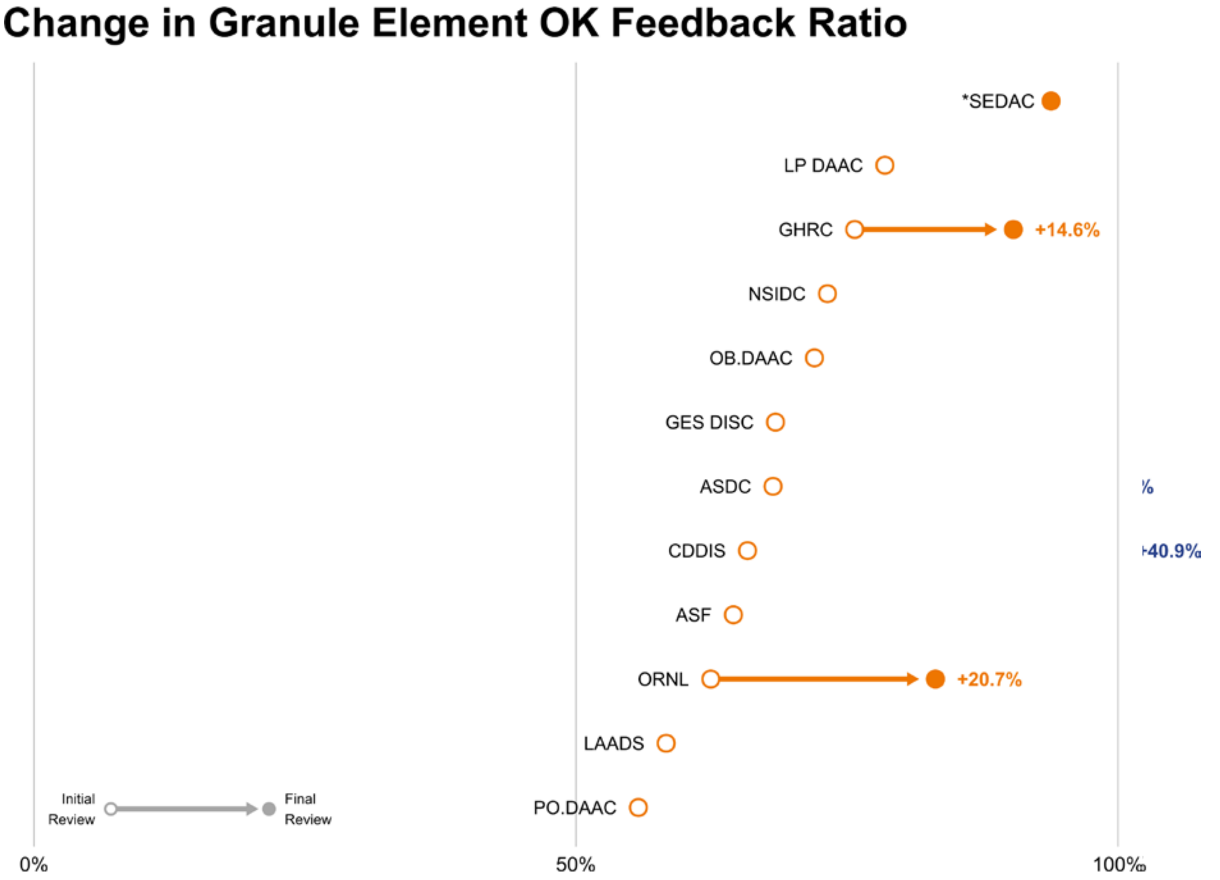
Process repeats to build community consensus around new policies and best practices

Phase I

- 1.5 years (mid 2016 – end 2017)
- Reviewed records from all 12 DAACs
- 1,961 collections reviewed
- GHRC, ASF, and CDDIS fully reviewed
- Supported CDDIS and SEDAC in the generation of brand new collection and granule metadata



Key Outcomes from Phase I


Change in Granule Element OK Feedback Ratio

✓ Reingested metadata is markedly improved at both the collection and granule levels

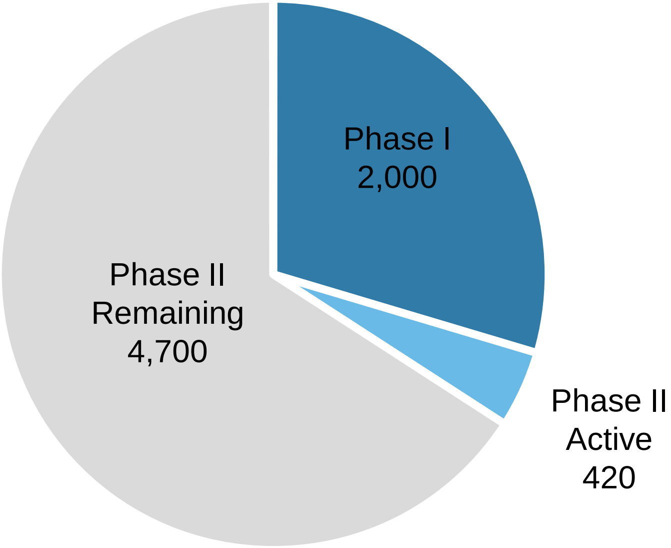
Phase II

- Remaining ARC reviews will transition to an online dashboard environment
  - Streamline dissemination of findings
  - Improve ARC/DAAC communication
  - Enable automated metric tracking

- Track DAAC improvements from Phase I

- Add clarity to existing UMM documentation and provide new reference resources for metadata authors
  - Work has just begun on building out a comprehensive Wiki space for UMM documentation

  https://wiki.earthdata.nasa.gov/display/CMR/UMM-C+Schema+Representation



Looking Forward

- ARC’s primary focus is delivering **actionable** feedback to the DAACs

- ARC is a one-off exercise; projected review completion is end of 2019

- Empower DAACs to provide more consistent and complete metadata by offering best practices and **improving documentation**
  - Easier to find
  - Easier to filter
  - Easier to consume

- UMM and associated mappings evolve

- When a DMSMM metric is output, how is utilized?
  - Is the intended audience a person? A machine?
  - How is it interpreted?
  - Should the metric be less than ideal, how does it become an actionable piece of information?

- ARC process is, to some extent, a manifestation of several of the rationales listed in CEOS WGISS DMSMM white paper
  - Thus, an implementation of the DMSMM would allow key elements of the ARC process to live beyond ARC itself
  - Important because the ARC process is not scalable in its current form

Questions

Adam Sisco
adam.sisco@nsstc.uah.edu

Kaylin Bugbee
kaylin.m.bugbee@nasa.gov

THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE