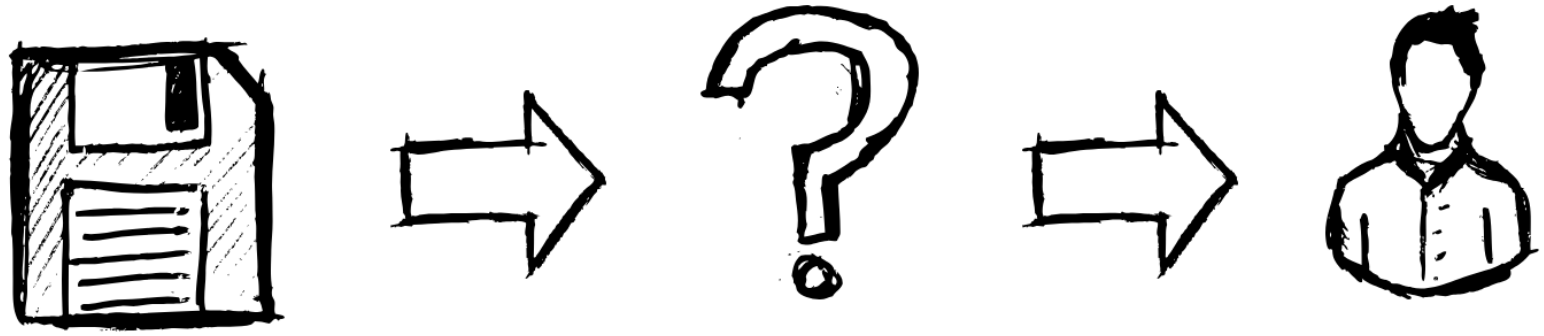
A Metadata Curation Approach to Improve the Discoverability and Accessibility of NASA Earth Science Data

Jeanné le Roux[1], Kaylin Bugbee[1], Valerie Dixon[2], Adam Sisco[1], Ingrid Garcia-Solera[1], Rahul Ramachandran[1]

(1) NASA MSFC IMPACT (2) NASA GSFC

What makes finding data possible?



Metadata

- Acts as a proxy for data
- Makes search possible
- Limits & focuses attention to the relevant information about a dataset

NASA's Common Metadata Repository (CMR)

- NASA's growing collection of data is archived and distributed by 12 Distributed Active Archive Centers (DAACs)
- The Common Metadata Repository (CMR) is a centralized repository which **stores all of NASA's Earth Science metadata**
- Nearly 7,000 NASA datasets (collections) and 370 million granules are described by metadata housed in the CMR

CMR metadata concepts

Collection
(Describes datasets)

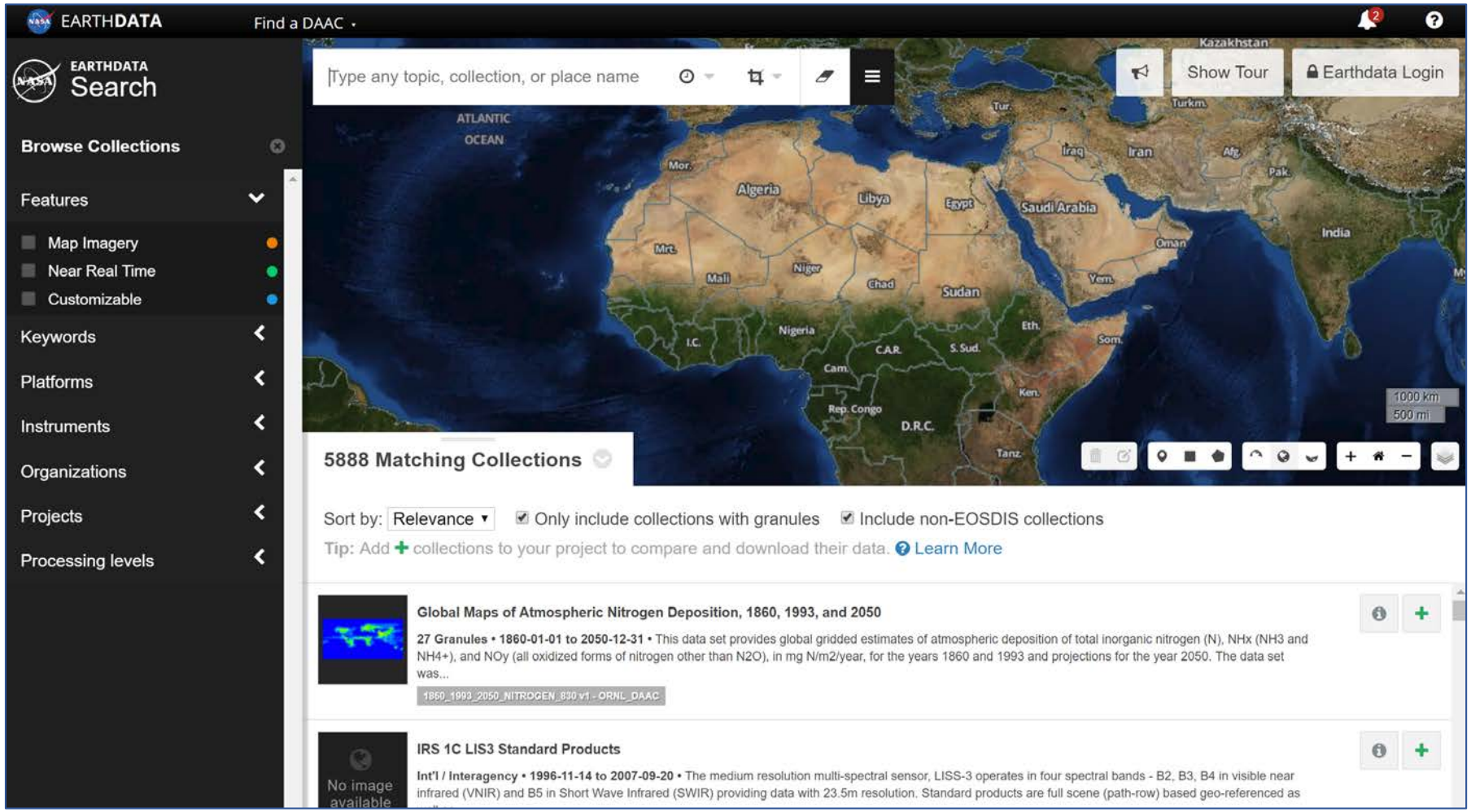
Granule
(File level metadata)

Concepts in development

Service
(Endpoints for sub-setting & transforming data)

Variable
(Describes science variables)

Earthdata Search



The Earthdata Search Client can be used to search across all metadata in the CMR. There is also a [CMR Search API].

- https://search.earthdata.nasa.gov

Curating NASA Metadata

- NASA recognizes the value of high quality metadata
  - Complete and correct metadata yields a better search experience

- Analysis and Review of CMR (ARC) Team:

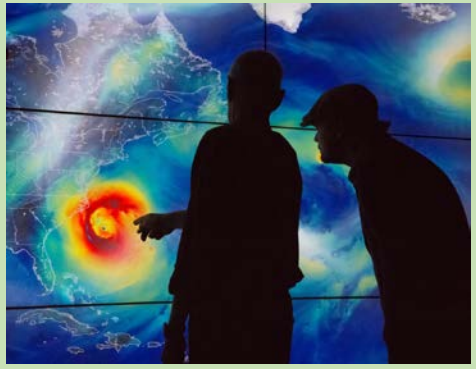
Team focused on curating NASA Earth Science metadata in the CMR for correctness, completeness, and consistency

- **Curating metadata → "inside-out" approach for easing discovery, accessibility, and usability of data**

ARC Curation Approach

- Ensuring metadata is correct, complete, and consistent makes data easier to discover

Curation is approached from a science user perspective.



Is the information provided useful to a broad user base?

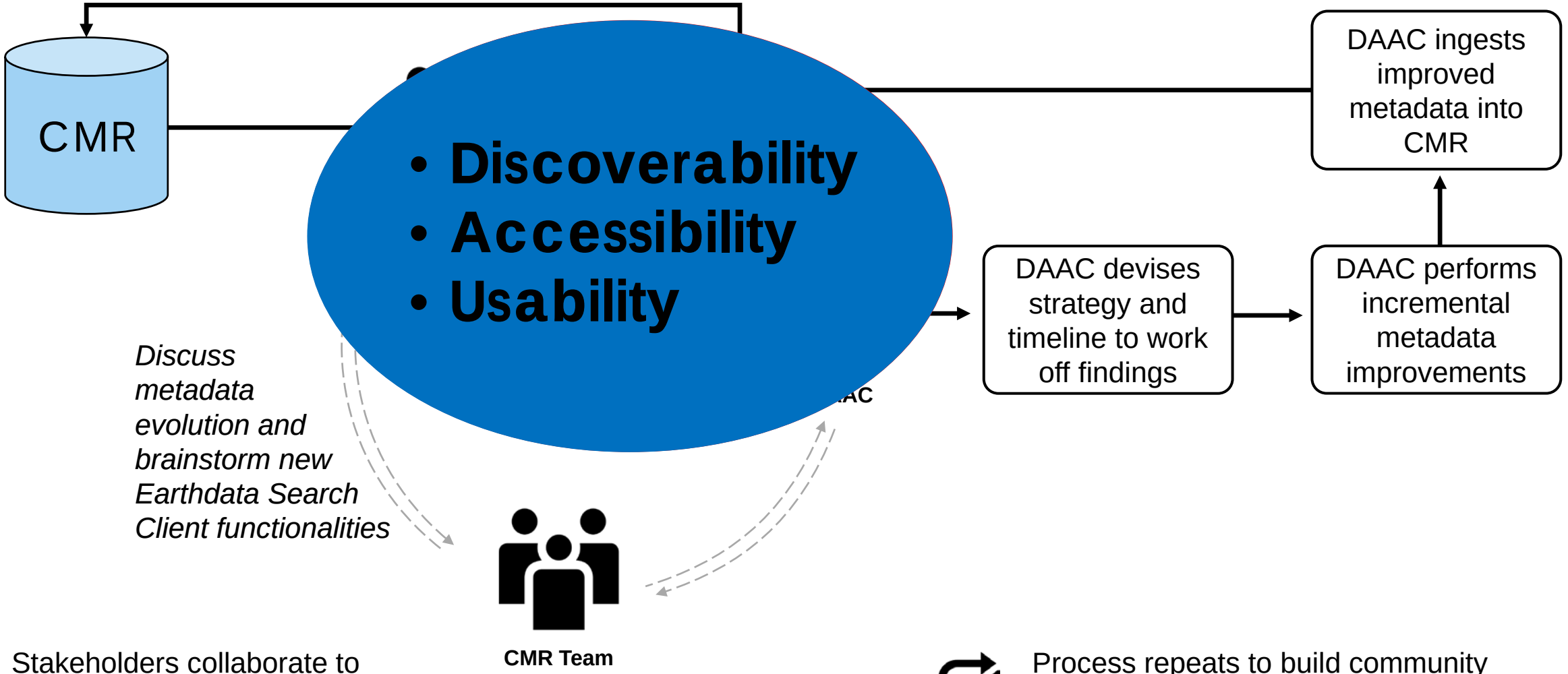
[Across disciplines within NASA]
[Across broader disciplines]

Human & Computer

While keeping in mind the needs and requirements of the CMR system.

[Schema compliance]
[Controlled vocabularies]

ARC Curation Process



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

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

Finding Data: Discoverability

Goal: Positive user experience

Want people to find the data they are looking for

Potential sources of frustration:

- No relevant search results
- Can't differentiate between similar datasets

Royal Navy official photographer, Tomlin, H W (Lt) [Public domain], via Wikimedia Commons

What do we look for in metadata?

Discoverability

- Correct temporal extent
- Correct spatial extent
- Abstract
  - Comprehensive
  - Useful to the science community but also approachable for a first time user of the data
- Does the metadata include appropriate keywords?
  - E.g. Science keywords, Platform and Instrument keywords, Data Center
- Correct processing level
- Is the title of the dataset descriptive?

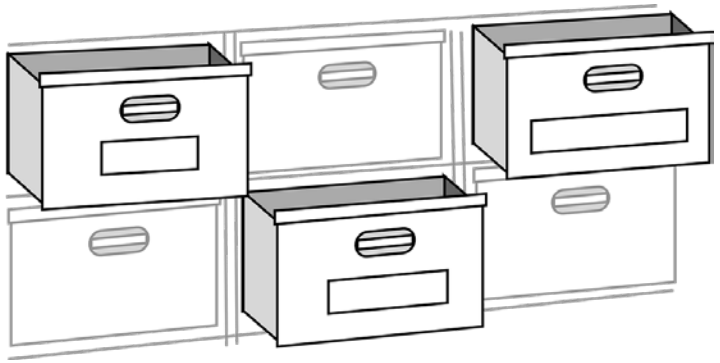
Once Data is Discovered: Accessibility

Goal: Positive user experience

Accessing data should be easy

Potential sources of frustration:

- Unclear how to access data
- No direct download option



What do we look for in metadata?

Accessibility

- Is data available via direct download?
- Is the correct data being served?
- Are users taken as directly to the described data as possible?
- Is all described data available?
- Is data access labeled correctly?

Once Data is Discovered: Usability

Goal: Positive user experience

Documentation should be readily available

Potential sources of frustration:

- Unclear what data contains (documentation lacking or too technical)
- Unclear where/ when/ how data were collected
- Unclear what tools needed to work with data

What do we look for in metadata?

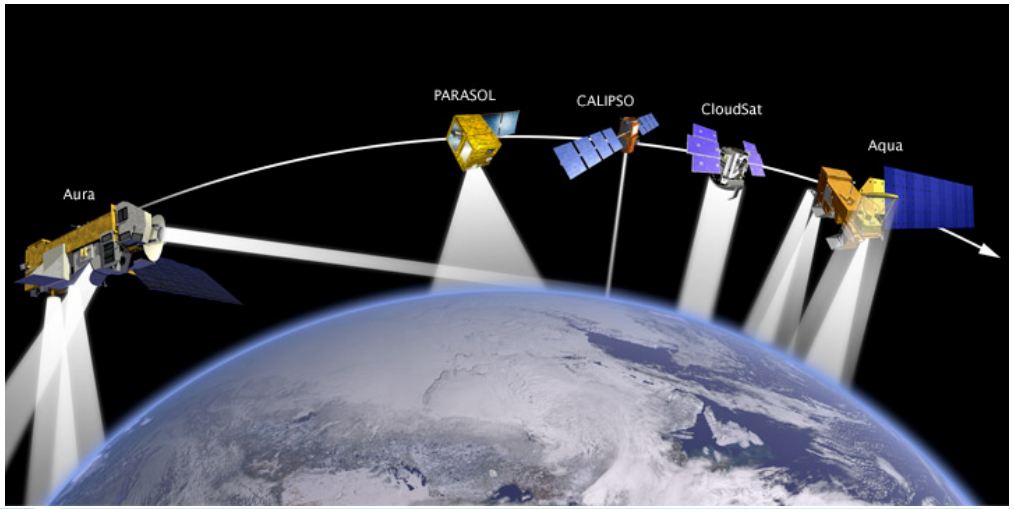
Usability

- Is all relevant online documentation provided?
  - Ensure online documentation is appropriately described/ labeled
  - User's guides, ATBDs, README files, FAQ pages, dataset landing page, etc.
- Data formats provided in metadata
- Links to relevant tools & services are provided, if applicable

In Summary

- ARC's metadata curation activities help ensure that NASA Earth science metadata is correct, complete, and consistent

- Checking for key items in the metadata helps facilitate discoverability, accessibility, and usability of the data



Looking Forward

Improvements to the metadata curation process:

- Comprehensive, centrally located best practices documentation under development on the Earthdata Wiki to help DAACs create/maintain metadata
- Development of more robust scripts to automate error detection as much as possible
- Improved support for bulk metadata updates

Session #: IN53C
Poster #: 0618

"Analysis and Review of NASA Earth Science Metadata: How Automation Plays a Role"

When? Today (Friday Dec 14) from 1:40 – 3:40

Questions?

Thank you!

Jeanné le Roux
jeanne.leroux@nsstc.uah.edu

Discovery: Beyond Metadata Curation

- Earthdata Search relevancy rankings
  - Effects discoverability
  - Inner workings of relevancy rankings is beyond the scope of the ARC project – ARC can ensure provided information is correct/relevant for the data
- Structured metadata for major search engines
  - E.g. Schema.org markup for Google
  - Beyond scope for this presentation
    - Focused on discovery within the CMR/ Earthdata Search

Accessibility: Beyond Metadata Curation

- Data organization structure
  - Sometimes data are organized in a way that is inconsistent with the metadata
  - Makes it difficult to point directly to described data
    - Different definitions of ‘dataset’
    - E.g. 1 File directory contains data from 4 different ‘datasets’, as specified in the collection level metadata record

Usability: Beyond Metadata Curation

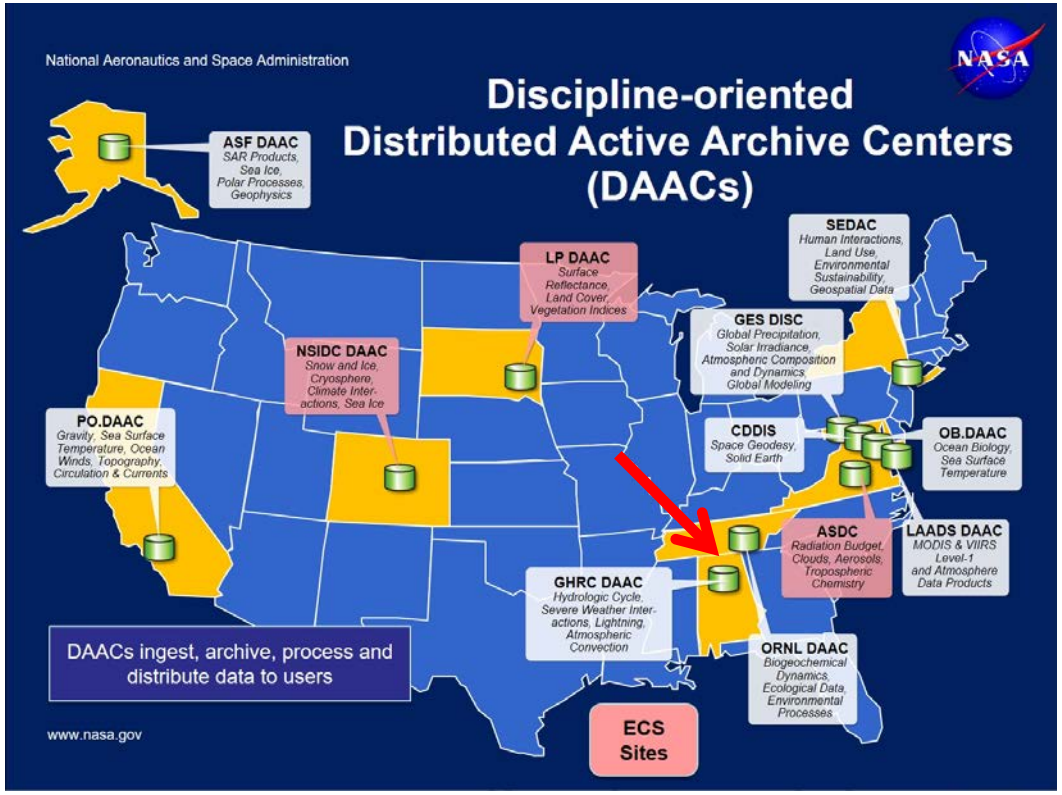
- Quality of dataset documentation
  - Lacking or highly technical user documentation is a barrier to data use
  - Improvements to dataset documentation is beyond ARC's scope
- Data format
  - 'Unfriendly,' outdated, or proprietary data formats may limit use of data
  - Especially if transformation services are not available

NASA Earth Science Data

NASA's Earth Observing System Data and Information System (EOSDIS)

Data is archived and distributed by 12 Distributed Active Archive Centers (DAACs)

Nearly 7,000 collections and 370 million granules are described by metadata housed in the **Common Metadata Repository (CMR)**



What is metadata curation?

Traditional curation

Digital curation

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

Metadata curation

Supports the research data lifecycle by ensuring the correctness, completeness and consistency of metadata

Digital Curation Center, Edinburgh, Scotland

Analysis and Review of CMR (ARC) Team

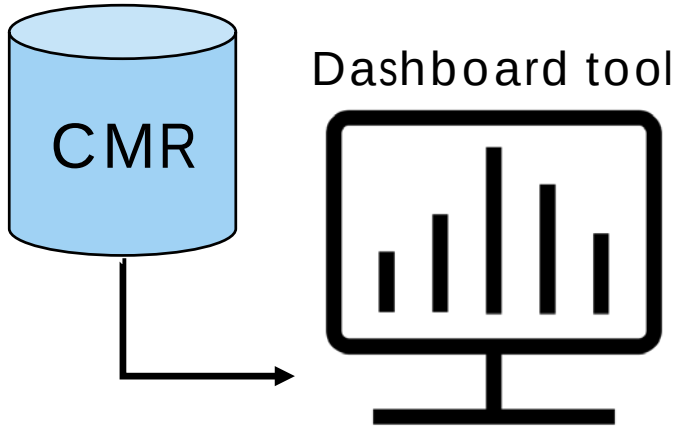
Team is comprised of Earth Science data and metadata specialists

Backgrounds in Earth system science, atmospheric science, space science, and remote sensing

Previous experience from the Climate Data Initiative (CDI)

- Review of 850 metadata records for quality and accessibility

ARC Curation Process



- Initial automated metadata checks
- Manual metadata review: metadata fields with issues are flagged
- Dashboard tracks improvement metrics

What do we look for in metadata?

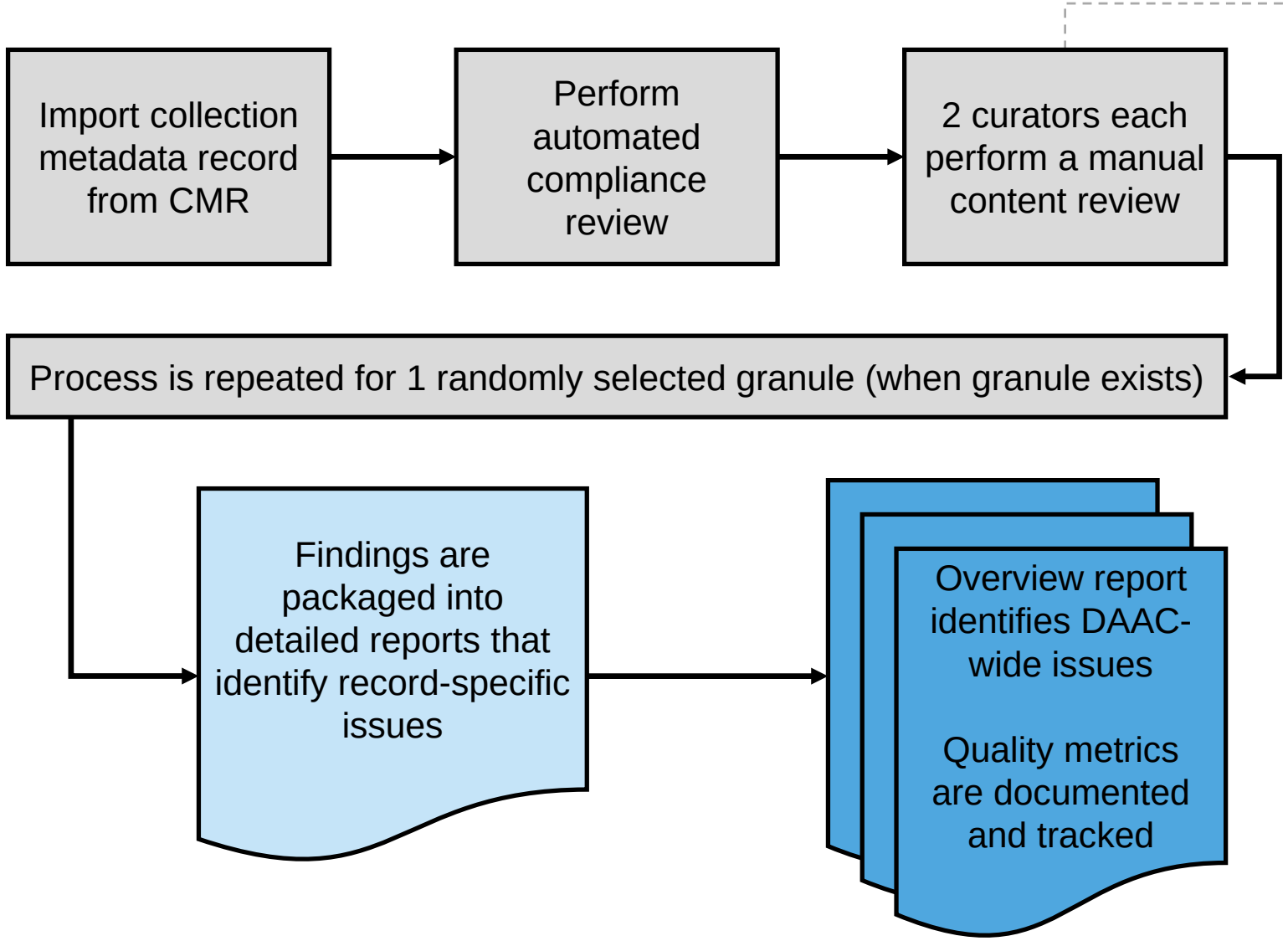
Compliance

- Required elements
- Controlled vocabulary
- Broken URLs
- DOIs

Content

- Accuracy
- Keyword relevance
- Missing information
- Consistency across records
- Comprehensibility

ARC Curation Process



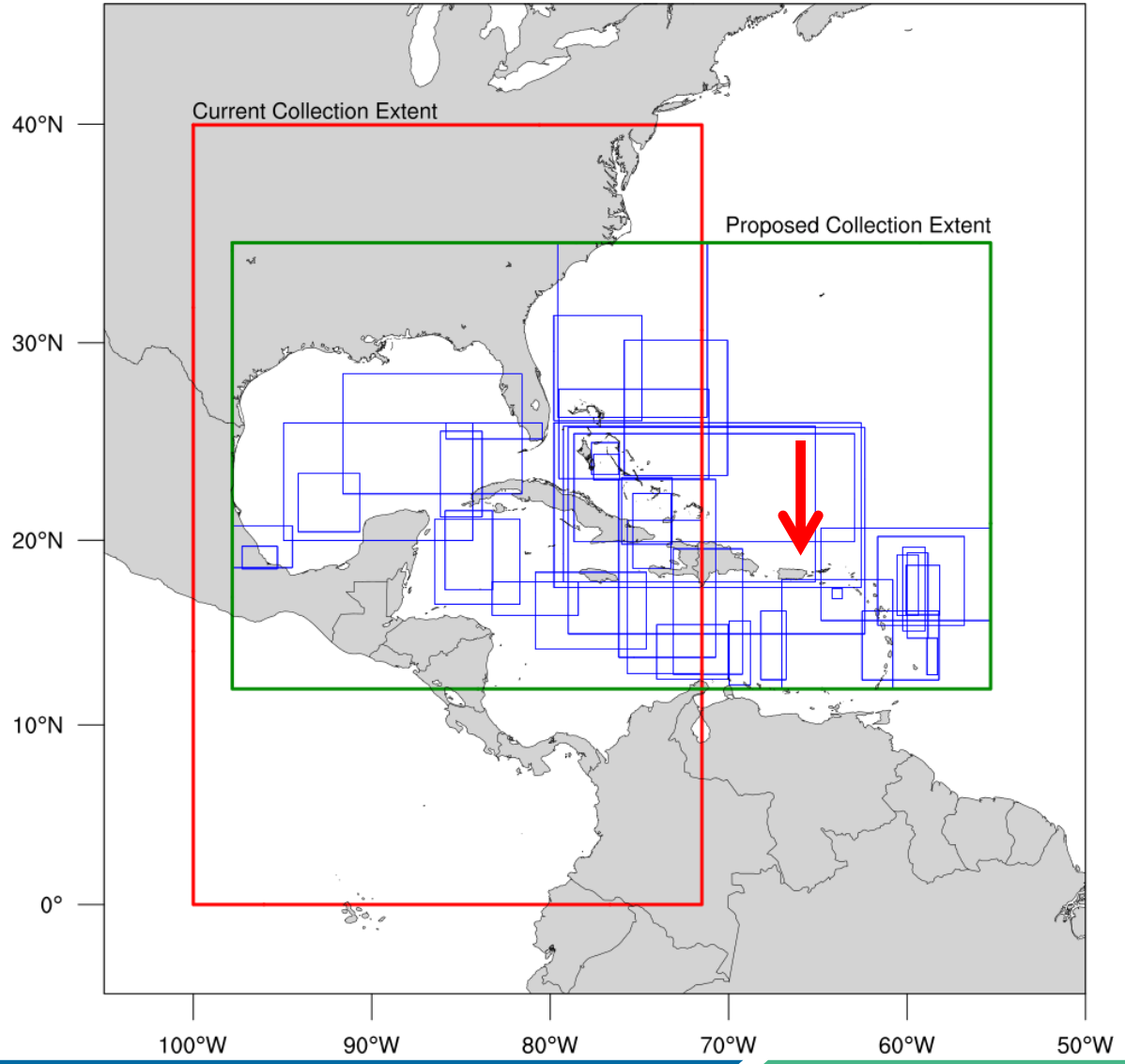
High	• Inaccurate, incomplete, or missing content • Broken URLs and invalid collection-granule relationships
Med	• Revisions of existing content • Addition of new information
Low	• Minor consistency issues

Priority classification scheme

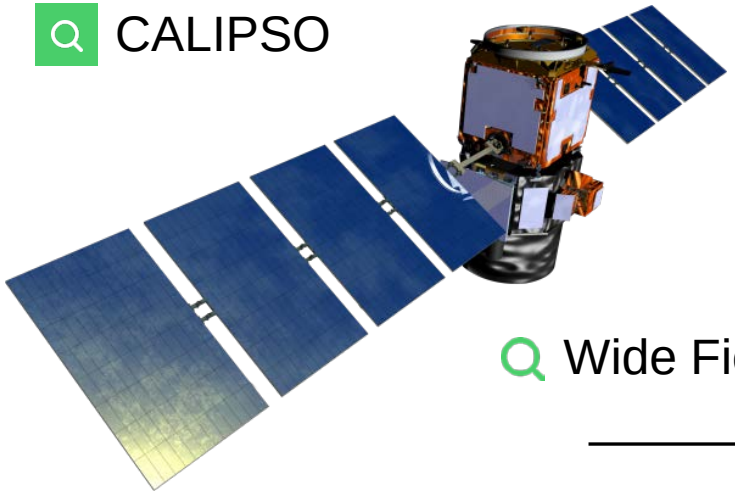
1. Assists DAAC in formulating a strategic plan to address findings
2. Used to track resolution of issues

Search and Discovery

Spatial Coverage



CALIPSO



- Wide Field Camera (WFC)
  - ⟶ 170K granules
- Imaging Infrared Radiometer (IIR)
  - ⟶ 449K granules
- Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP)
  - ⟶ 1 granule

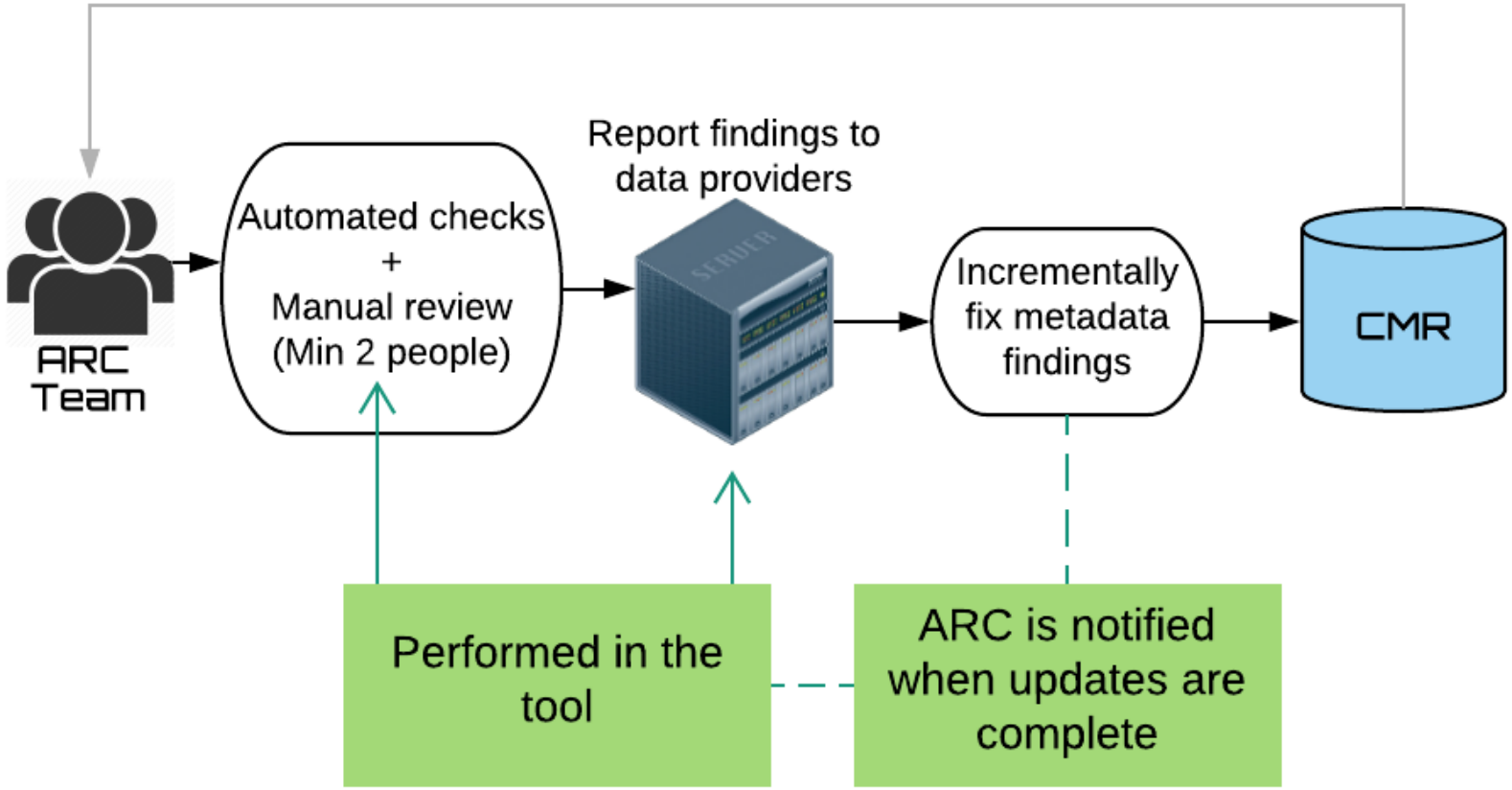
LIDAR 2M granules

Metadata Example: JSON Snippet


```
  "ShortName": "CIESIN_SEDAC_GISS_CROPCLIM_DB",
  "Abstract": "Potential Impacts of Climate Change on World Food Supply: Datasets from a Major Crop Modeling Study contain projected country and regional changes in grain crop yields due to global climate change. Equilibrium and transient scenarios output from General Circulation Models (GCMs) with three levels of farmer adaptations to climate change were utilized to generate crop yield estimates of wheat, rice, coarse grains (barley and maize), and protein feed (soybean) at 125 agricultural sites representing major world agricultural regions. Projected yields at the agricultural sites were aggregated to major trading regions, and fed into the Basic Linked Systems (BLS) global trade model to produce country and regional estimates of potential price increases, food shortages, and risk of hunger. These datasets are produced by the Goddard Institute for Space Studies (GISS) and are distributed by the Columbia University Center for International Earth Science Information Network (CIESIN).",
  "DirectoryNames": [
    {
      "ShortName": "USA/CIESIN"
    }
  ],
  "Purpose": "To provide an assessment of potential climate change impacts on world crop production, including quantitative estimates of yield changes of major food.",
  "PublicationReferences": [
    {
      "PublicationDate": "2009-05-01T00:00:00.000Z",
      "Title": "Effects of Climate change on Global Food Production Under SRES Emissions and Socio-economic Scenarios",
      "DOI": {
        "DOI": "10.7927/H4JM27JZ"
      },
      "OnlineResource": {
        "Linkage": "https://doi.org/10.7927/H4JM27JZ"
      },
      "Publisher": "NASA Socioeconomic Data and Applications Center (SEDAC)",
      "Author": "Iglesias, A. and C. Rosensweig",
      "PublicationPlace": "Palisades, NY"
    }
  ],
  "DOI": {
    "DOI": "10.7927/H43R0QR1"
  },
  "RelatedUrls": [
    {
      "URLContentType": "VisualizationURL",
      "Type": "GET RELATED VISUALIZATION",
      "URL": "http://sedac.ciesin.columbia.edu/downloads/maps/crop-climate/crop-climate-potential-impacts-world-food-supply/sedac-logo.jpg"
    },
```

What is the Metadata Curation Dashboard?

- Tool used for conducting metadata quality reviews
- Records findings for reporting and metrics tracking

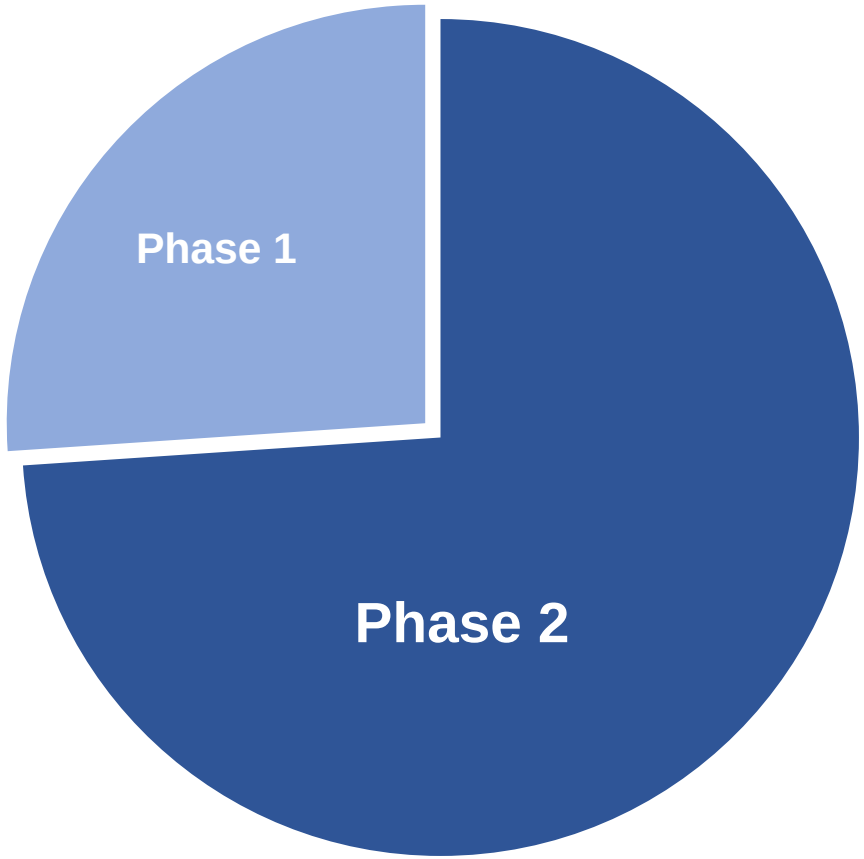


Phase I Metrics

Reviewed over 25% of collection level records in CMR

- Records from all 12 data centers reviewed

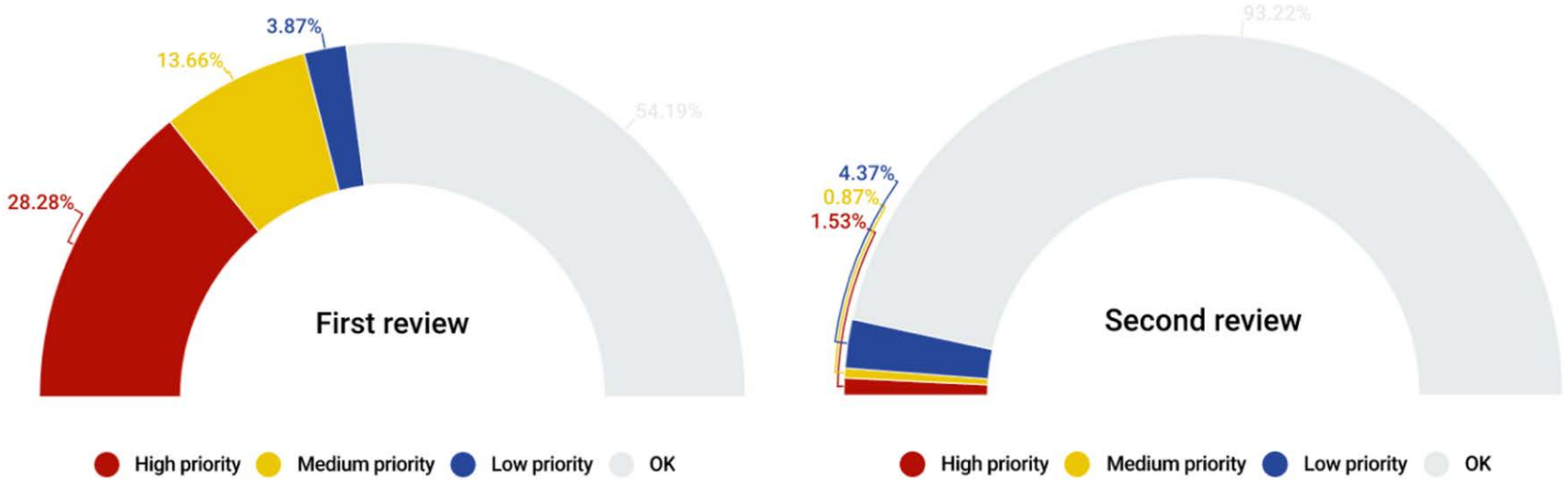
Supported two data centers in the generation of brand new collection and granule metadata



1,959 collections reviewed

Key Outcomes from Phase I

ORNL



SEDAC