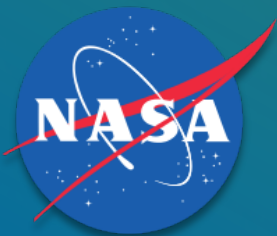


Evolving NASA's Data and Information Systems for Earth Science

Dr. Rahul Ramachandran

Manager | Inter-Agency Implementation and Advanced Concepts Team (IMPACT)
Senior Research Scientist | Earth Science Branch (ST11)
Marshall Space Flight Center/NASA
Huntsville, Alabama 35812, USA

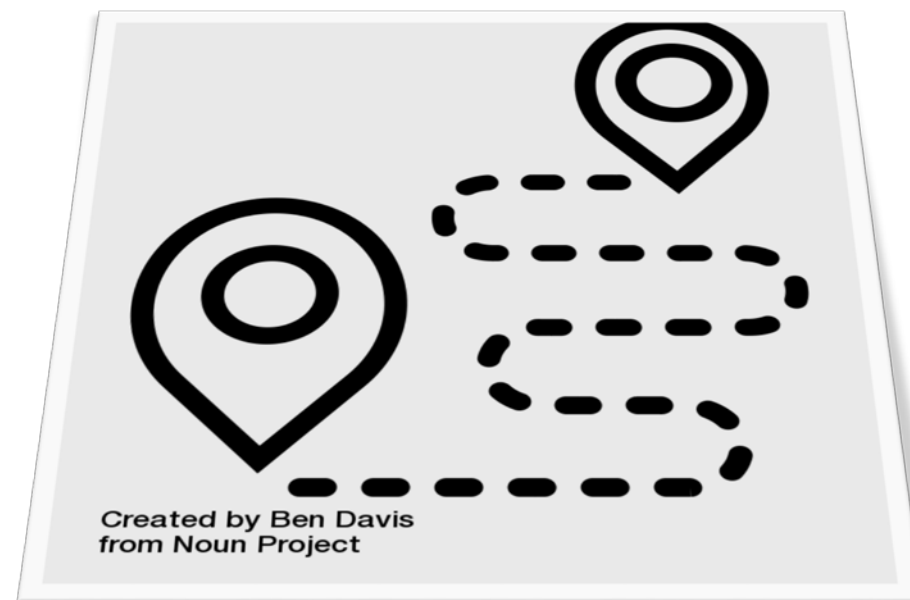


HPC User Forum, April 1-3 2019 Santa Fe, NM



Presentation Roadmap

- Overview
 - What, Why and How?
- Data Discovery
 - Metadata
- Data Use
 - Cloud Infrastructure
 - Analytics



*roadmap by Ben Davis from the Noun Project

An aerial photograph of a coastline is partially visible in the top-left corner, showing a dark, rocky shore meeting the sea. The rest of the image is covered by a large, semi-transparent green overlay that has a subtle, wavy texture. The word "Overview" is centered in the middle of this green area in a white, sans-serif font. Below the text, a thin white horizontal line spans across the width of the text.

Overview

Earth Science – NASA’s Strategic Goal

This ability to *observe our planet comprehensively* matters to each of us, on a daily level. Earth information—for use in Internet maps, daily weather forecasts, land use planning, transportation efficiency, and agricultural productivity, to name a few—is central to our lives, providing substantial contributions to our economies, our national security, and our personal safety. It helps ensure we are a thriving society. - NRC, 2018

NASA’s Strategic Goal 1.1:
“Understand The Sun, Earth, Solar System, And Universe.”



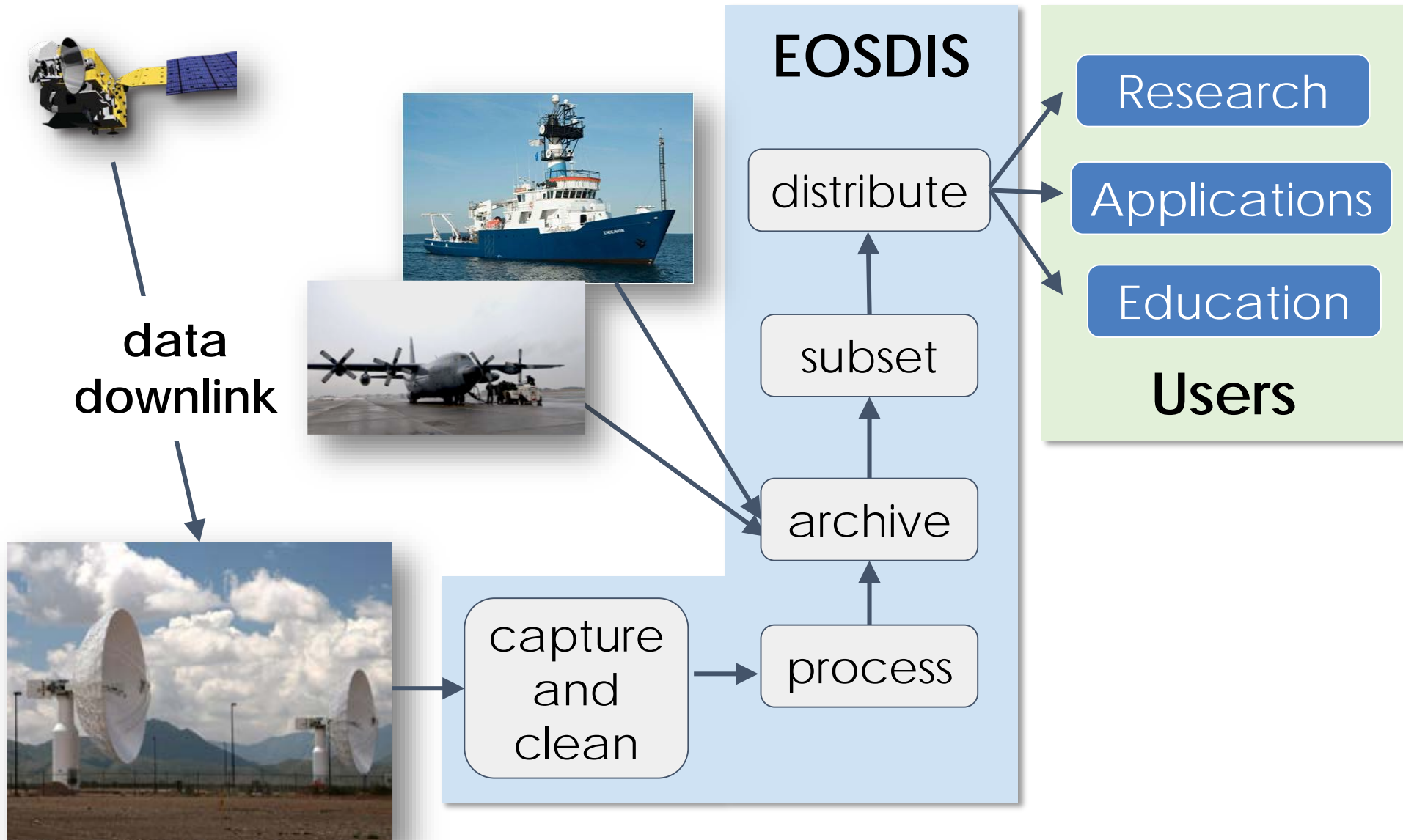
Earth Science Data System Program

The Earth Science Data System Program is an essential component of the Earth Science Division and is responsible for:

- Actively managing NASA's Earth science data (Satellite, Airborne, and Field).
- Developing unique data system capabilities optimized to support rigorous science investigations and interdisciplinary research.
- Processing (and reprocessing) instrument data to create high quality long-term Earth science data records.
- Upholding NASA's policy of full and open sharing of all data, tools, and ancillary information for all users.
- Engaging members of the Earth science community in the evolution of data systems.

The Earth Science Data and Information System (ESDIS) project at GSFC maintains and operates a data and information system for NASA's Science Mission Directorate (SMD) and its Earth Science Division (ESD) to support multidisciplinary research in Earth science and public data access.

Earth Observing System Data and Information System (EOSDIS)



Earth Observing System Data and Information System (EOSDIS)

EOSDIS is managed by the Earth Science Data and Information System (ESDIS) Project at GSFC and includes the following major core components:

Science Investigator-led Processing Systems (SIPS)

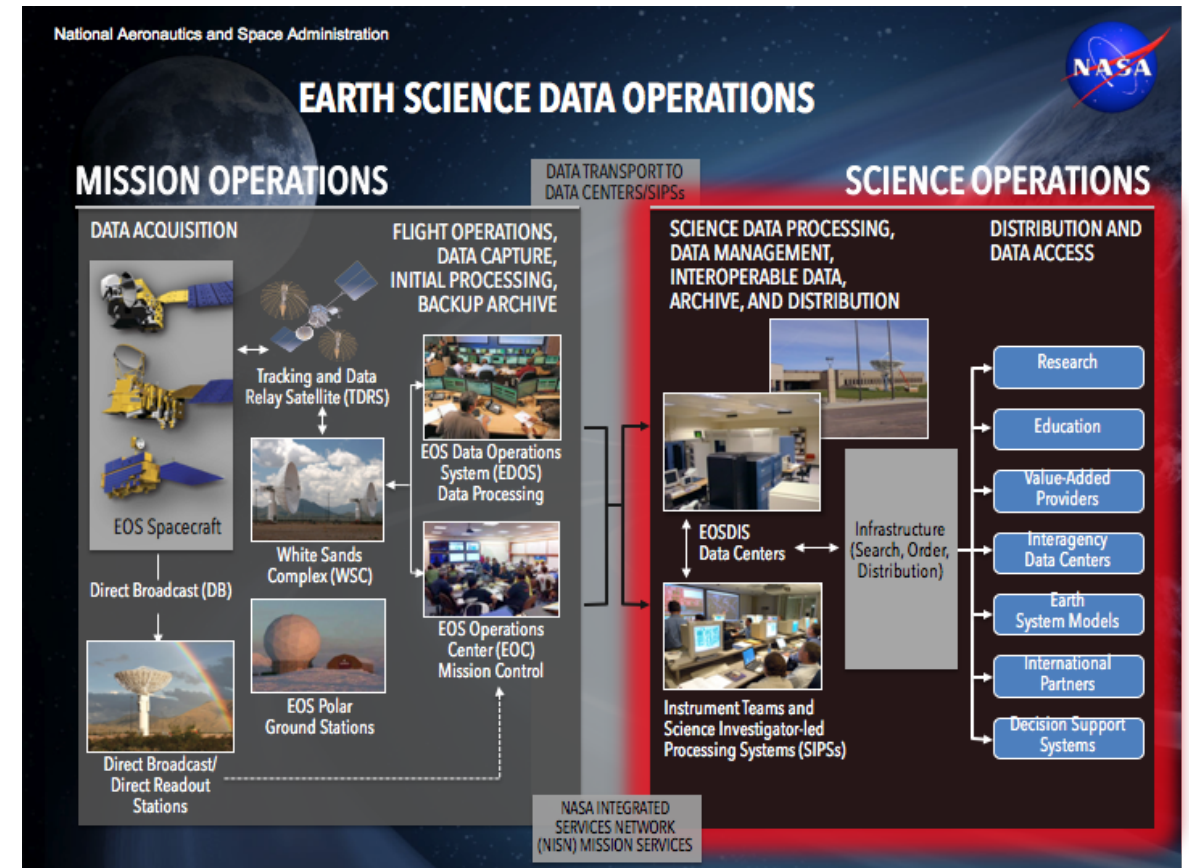
- Perform forward processing of standard data products and reprocess data to incorporate algorithm improvements

Distributed Active Archive Centers (DAACs)

- Co-located with centers of science discipline expertise; archive and distribute standard data products produced by the SIPS and others

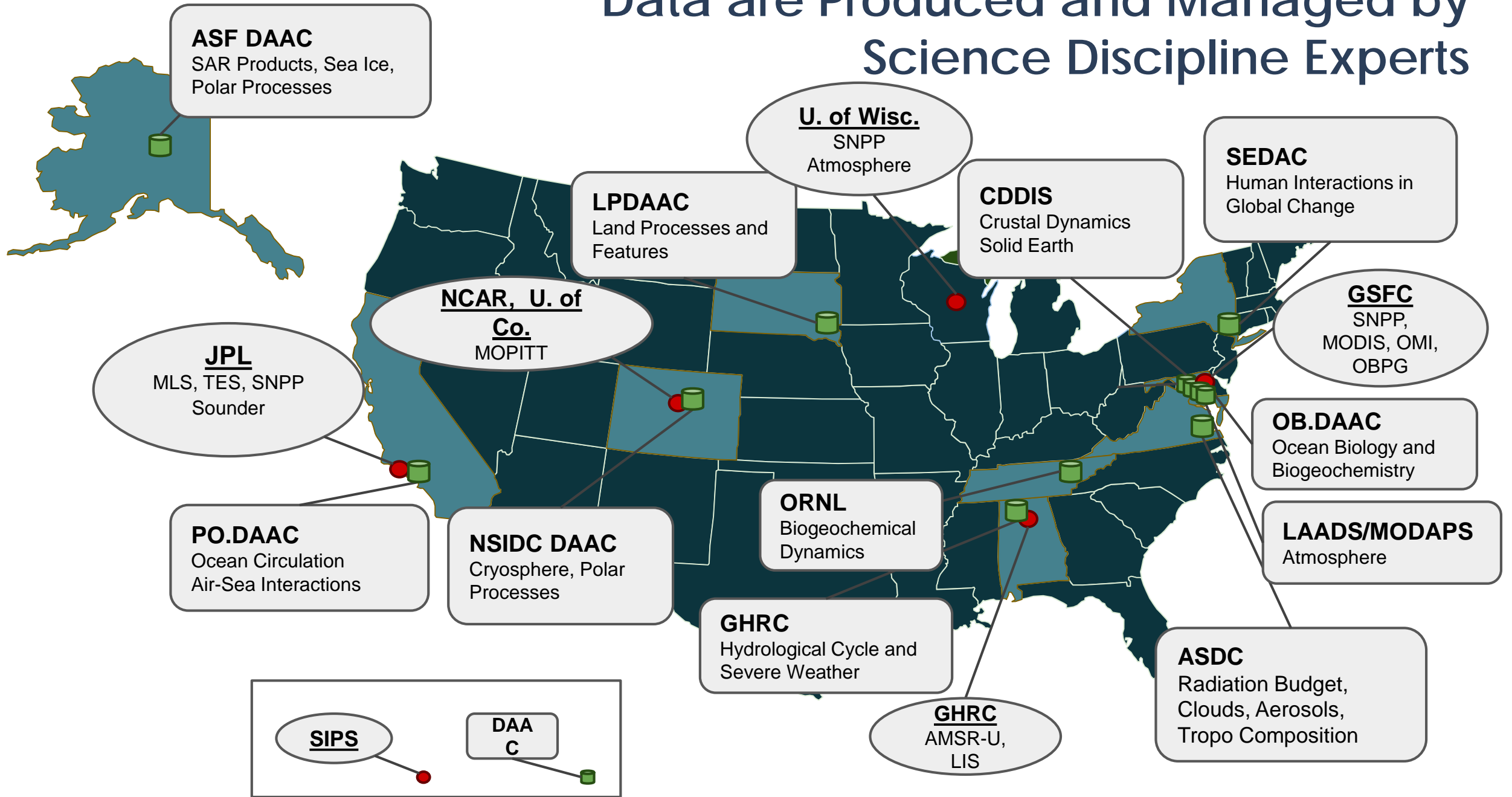
Earthdata and Core Services

- Allows users to search, discover, visualize, refine, and access NASA Earth Observation data. Includes networking and security



*Red highlight indicates EOSDIS boundary.

Data are Produced and Managed by Science Discipline Experts



Extensive Data Collection

Started in the 1990s, EOSDIS today has 11,000+ data types (collections)

- Cover & Usage
- Surface temperature
- Soil moisture
- Surface topography

Land



- Surface temperature
- Surface wind fields & heat flux
- Surface topography
- Ocean color

Ocean



- Winds & Precipitation
- Aerosols & Clouds
- Temperature & Humidity
- Solar radiation

Atmosphere



- Population & Land Use
- Human & Environmental Health
- Ecosystems

Human Dimensions



- Sea/Land Ice & Snow Cover

Cryosphere



Data Sources

Type	Example Missions
Satellite/on-orbit Missions	Terra, Aqua, Aura, Suomi-NPP, SORCE, GPM, GRACE, CloudSat, CALIPSO, etc.
Airborne Missions	IceBridge, Earth Ventures (5+ missions), UAVSAR, etc.
In Situ Measurement Missions	Field campaigns on land (e.g., LBA-ECO) and in the ocean (e.g., SPURS)
Applications support	Near-real time creation and distribution of selected products for applications communities
Earth Science Research support	Research products from efforts like MEaSURES. This also includes data from older, heritage missions (prior to EOS Program) that the DAACs rescue – e.g., Nimbus, SeaSat

NASA Earth Science Data and Information Policy

In effect since the early 1990s, **full and open sharing of data** from satellites, sub-orbital platforms and field campaigns with all users as soon as such data become available.

- **No period of exclusive access.** Following a post-launch checkout period, all data will be made available to the user community. Any variation in access will result solely from user capability, equipment, and connectivity.
- **Make available** all NASA-generated standard products along with the **source code for algorithm software, coefficients, and ancillary data** used to generate these products.
- **Non-discriminatory data access** so that all users will be treated equally.
- **Ensure that all data required for Earth system science research are archived.** Include easily accessible information about data holdings - quality assessments, supporting relevant information, and guidance for locating and obtaining data.
- **Interagency cooperation** - sharing of data from satellites and other sources, mutual validation and calibration data, and consolidation of duplicative capabilities and functions.
- **Collect metrics** to assess efficacy of data systems and services, and assess **user satisfaction**.

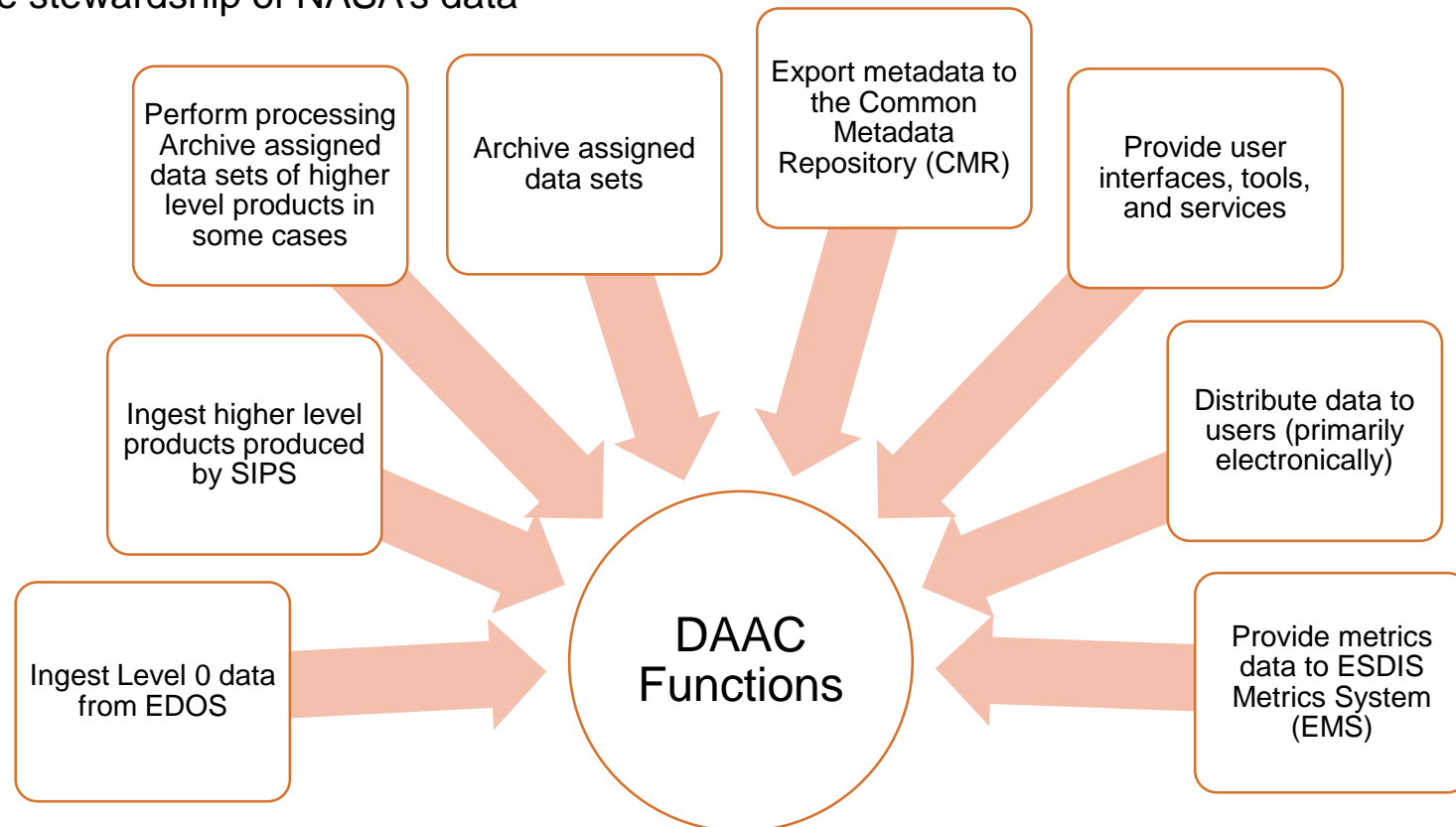
Science Investigator-led Processing Systems (SIPS) Requirements

- Perform forward processing and produce standard data products from mission instrument data; includes:
 - Generation of Level 1, Level 2 and/or Level 3 granules
 - Associated granule-level metadata
 - Associated browse products (as appropriate)
- Deliver all standard data products to the assigned DAAC so they can be available within 48 hours of the observation.
- Produce or enable production of near-real time data products to meet delivery (latency) goals.
- Reprocess standard data products to reflect algorithm improvements and to ensure consistent time series.
- Deliver documentation to the assigned DAAC for distribution.
- Assist the DAAC with information related to scientific content, format and product generation history of SIPS products.
- Deliver each version of data production source code used in production of the Standard Products.
- Participate with DAAC in collection of items for long-term preservation

Role of DAACs

DAACs were selected and established based on the Earth Science discipline expertise and heritage of their host organizations.

- Provide unique support and expert services to their user communities
- Provide data and services to the research community for comprehensive, cross-discipline studies needed to understand Earth as an interrelated system
- Ensure safe stewardship of NASA's data




Procedures for Archiving

No matter what type of data (on orbit, aircraft, in situ, etc.) DAAC Staff follow these procedures for handling datasets

Planning	<ul style="list-style-type: none">• Collaborate with mission teams, data producers, and ESDIS to develop Interproject Agreements (IPAs), Interface Control Documents (ICDs), and Operations Agreements (OAs)• Support data producers with the creation of Data Management Plans (DMPs)
Acquire and/or Produce Data	<ul style="list-style-type: none">• Advise data producers on data formats, structure, and delivery methods• Establish automated processes for the transfer of data into DAAC data systems• Develop or integrate, test, verify, and run data production code (for applicable data)
Preserve Data	<ul style="list-style-type: none">• Ensure redundant online disk and tape data storage• Establish file management (e.g. duplicate file detection) and file integrity (e.g. checksum verification)
Describe Data	<ul style="list-style-type: none">• Create collection-level metadata and, when necessary, employ software to extract file-level metadata for the purposes of preservation, discovery and usage• Export metadata to NASA's CMR for inter-mission and inter-sensor data discovery• Develop user documentation and supplemental information• Create DOIs and data citations for proper attribution of data and data creators
Distribute Data	<ul style="list-style-type: none">• Provide discovery through the DAAC Web site, with direct HTTPS access• Support automated data transfer to users through subscriptions and APIs• Facilitate search, visualization, and customization through NASA Earthdata Search• Develop specialized portals and data services per mission and user needs
Support Data	<ul style="list-style-type: none">• Assist user communities with the selection and usage of data• Create "How To" guides, FAQs, and other resources to address specific user needs• Work with user communities to identify needed improvements for data and tools• Provide outreach and education to broaden the user community

Levels of Service

To be cost efficient, not every dataset gets the same level of service

BASIC 

Includes all **Basic** services:

DATA


- File sizes, checksums, and number of files have been verified.
- File names are descriptive and consistent.
- File format and structure are appropriate for expected data use.
- Key metadata—geospatial, temporal, and science variable information—are provided and well-defined.
- Data can be accessed via all supported methods.
- Data are backed up and versioned.

DOCUMENTATION

- Data discovery and usage metadata are available.
- Links are provided to supporting documentation describing data content and methodologies.
- Data set landing page and data citation, including Digital Object Identifier (DOI), are available.

USER SUPPORT

- Assistance with data access.
- Assistance with basic data usage questions and referral to external documentation or data provider for more complex questions.

STANDARD 

Includes all **Basic** services **plus**:

DATA


- Data usability has been verified in select data analysis tools.

DOCUMENTATION

- User guide is provided with the following content: detailed descriptions of science variables, geospatial and temporal information, and data quality.

USER SUPPORT

- Assistance with data usage questions.
- Guidance on use of data in select data analysis tools.

COMPREHENSIVE 

Includes all **Standard** services **plus**:

DATA

- Data customization services—subsetting, reformatting, and/or reprojection—are available for select data.

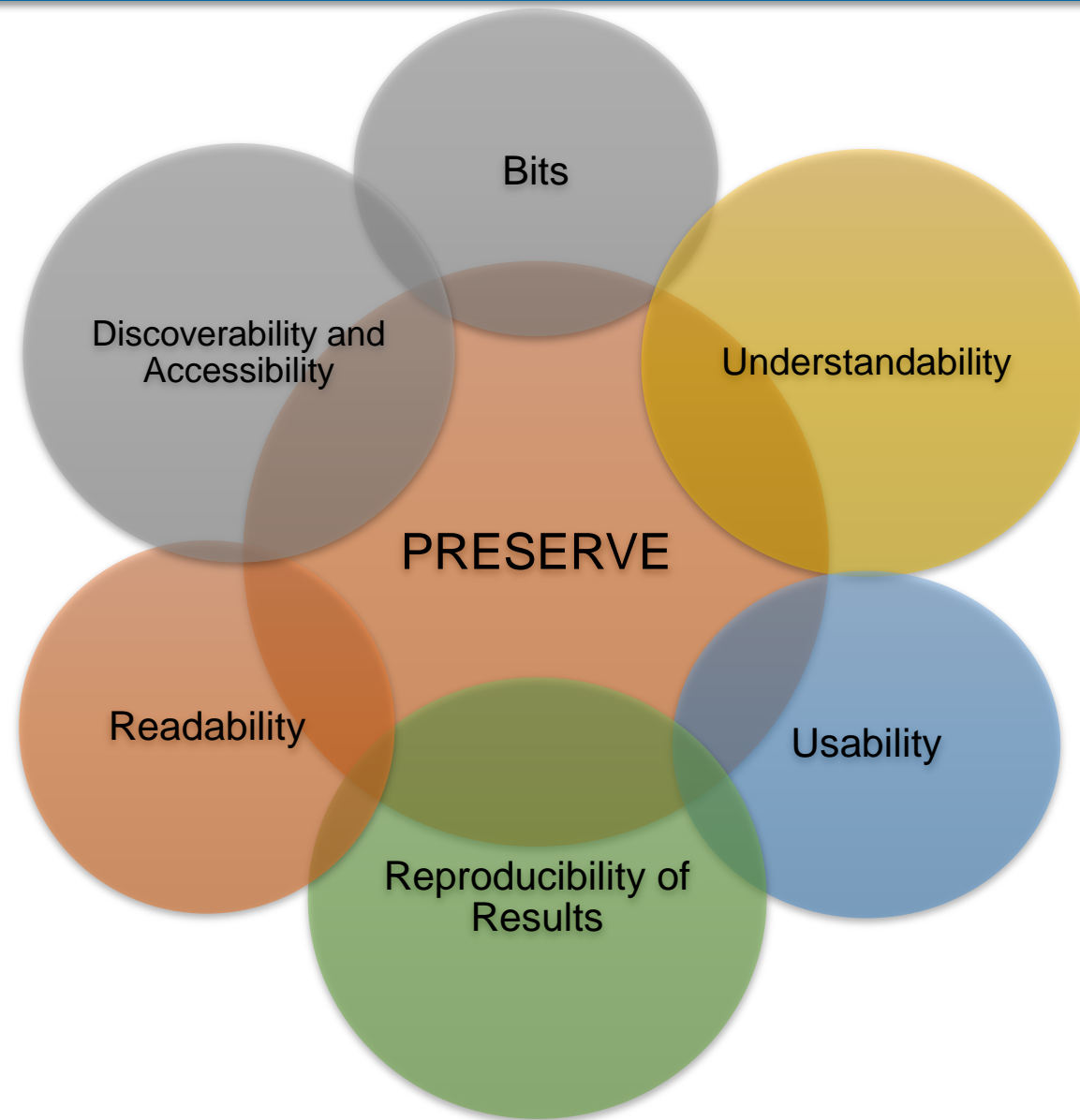
DOCUMENTATION

- Comprehensive user guide is provided with the following content: detailed descriptions of science variables, geospatial and temporal information, data quality, and data methodologies.

USER SUPPORT

- Assistance with complex data usage and methodologies questions.
- Guidance on use of data customization services for select data.

Preservation involves ensuring long-term protection of...



NASA's Preservation Content Specification for Earth Science Data

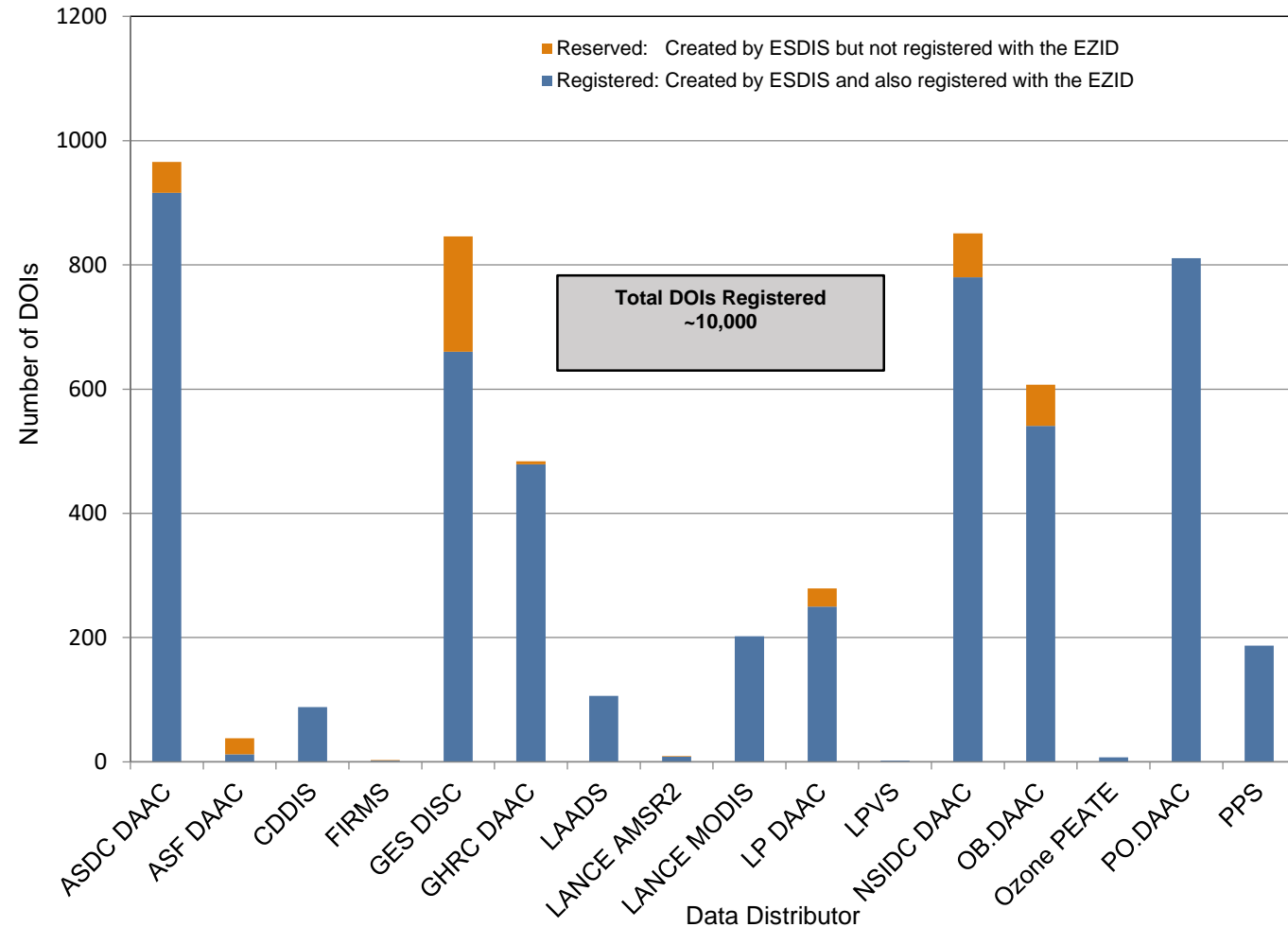
ESDIS Project has supported this as an addition to the ISO – 19165 – “Geographic Information - Preservation of digital data and metadata” We submitted 19165-2 as specific to Earth Observation Data

1. Preflight/Pre-Operations: Instrument/Sensor characteristics including pre-flight/pre-operations performance measurements; calibration method; radiometric and spectral response; noise characteristics; detector offsets
2. Science Data Products: Raw instrument data, Level 0 through Level 4 data products and associated metadata
3. Science Data Product Documentation: Structure and format with definitions of all parameters and metadata fields; algorithm theoretical basis; processing history and product version history; quality assessment information
4. Mission Data Calibration: Instrument/sensor calibration method (in operation) and data; calibration software used to generate lookup tables; instrument and platform events and maneuvers
5. Science Data Product Software: Product generation software and software documentation
6. Science Data Product Algorithm Input: Any ancillary data or other data sets used in generation or calibration of the data or derived product; ancillary data description and documentation
7. Science Data Product Validation: Records, publications and data sets
8. Science Data Software Tools: product access (reader) tools.

Digital Object Identifiers/Citations

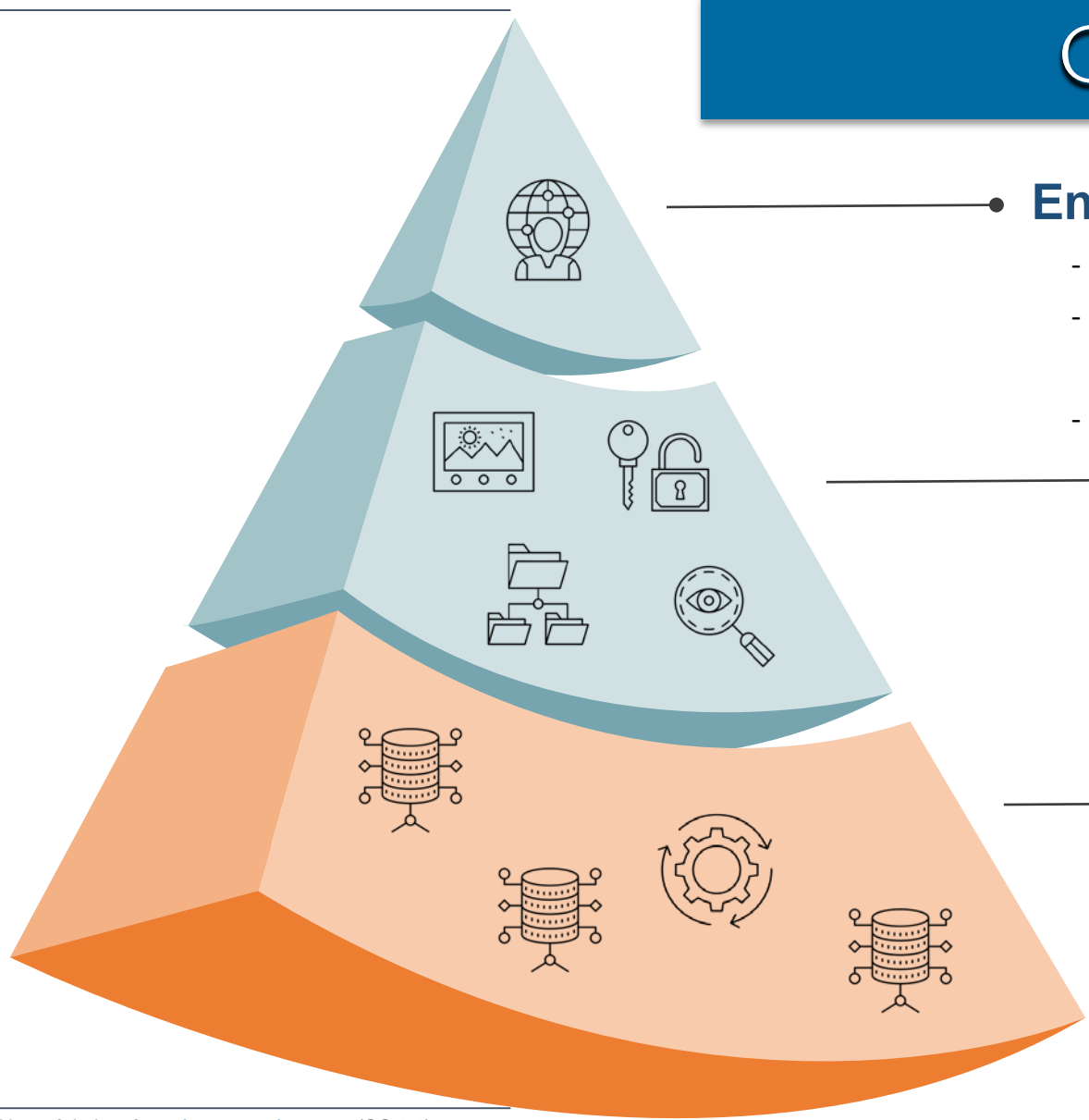
Status of Digital Object Identifiers created with the ESDIS
by Data Providers as of July 31, 2018

- Authors using data in publications should cite and reference the datasets properly
- DOIs are assigned to EOSDIS datasets by the ESDIS Project
- Associated with each DOI is a common landing page that provides details about datasets and shows how to cite them



Core Services

Metrics!



End User Web Clients

- **Earthdata.nasa.gov**
- **Earthdata Search:** data access/discovery*
- **Worldview:** imagery*

Open Service APIs

- **CMR:** Metadata Catalog (Search Engine)*
- User Login
- **GIBS:** Global Imagery Browse Services*

Earth Science Data Holdings

- Open APIs
- Free Data Download
- DAAC specific tools

EOSDIS core tools

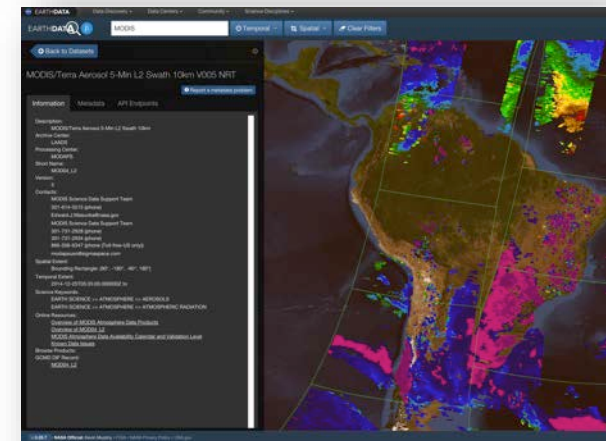
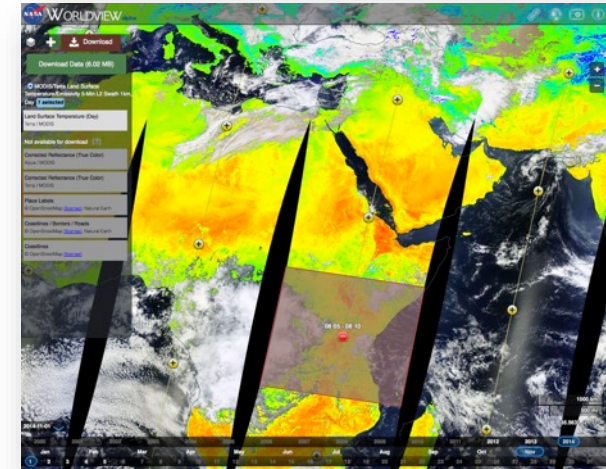
Federated resources

*open source software

all by Dinosoft Labes from thenounproject.com (CC 3.0)

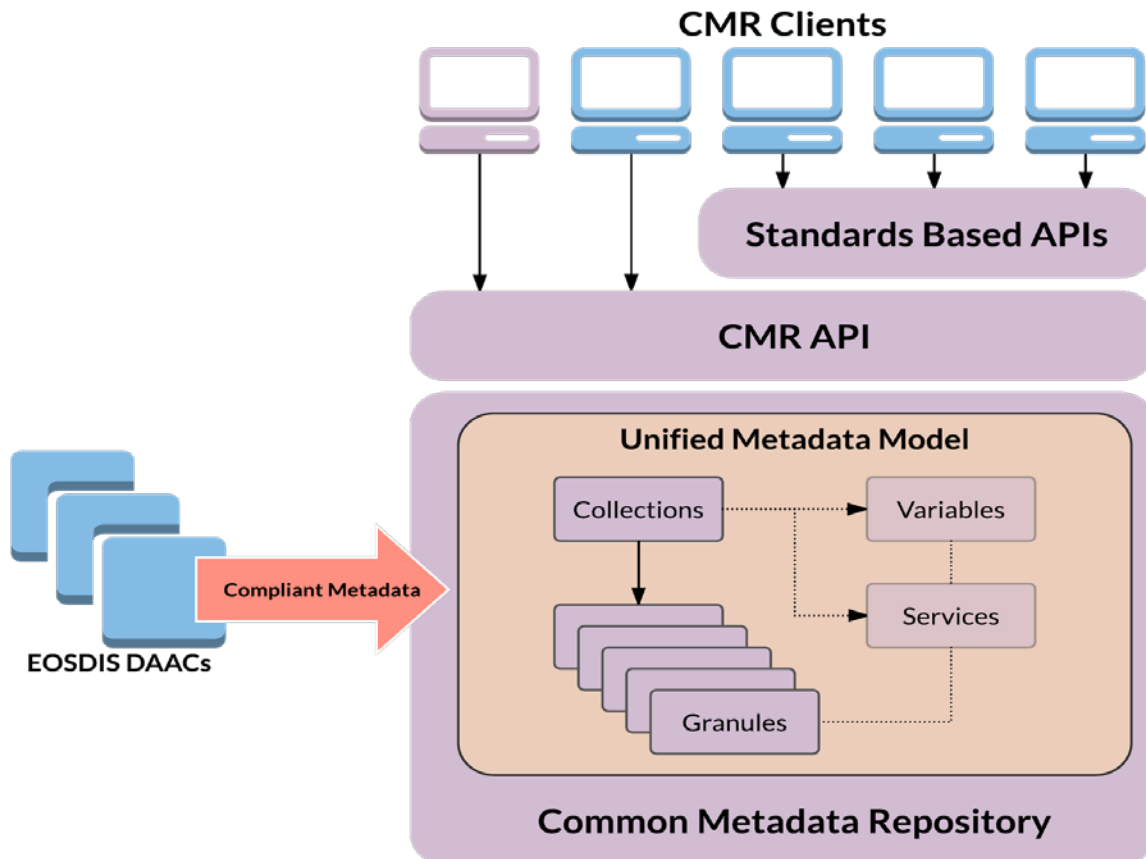
Data Access Centralized Reusable Capabilities

- **Earthdata:** The EOSDIS website <https://earthdata.nasa.gov> will increase visibility to the interdisciplinary use of data and demonstrate how data are used.
- High Performance Data Search and Discovery
 - **Common Metadata Repository (CMR):** Provide sub-second search and discovery services across the Sentinel and other EOSDIS holdings.
 - **Earthdata Search Client:** Data search and order tool <https://search.earthdata.nasa.gov>
- Imagery and Data Visualization Tools
 - **Global Imagery Browse Services (GIBS):** full resolution imagery in a community standards-based set of imagery services
 - **Worldview:** highly responsive interface to explore GIBS imagery and download the underlying data granules <https://earthdata.nasa.gov/labs/worldview/>
 - **Giovanni:** Quick-start exploratory data visualization and analysis tool
- **Near Real-Time Capabilities:** Provided by “**LANCE**” (Land Atmosphere Near real-time Capability for EOS) which produces products within < 3 hours of observation. Near real-time capabilities are co-located with the standard science production facilities.
- **EOSDIS Metrics System (EMS):** collects and reports on data ingest, archive, and distribution metrics across EOSDIS
- **Earthdata Infrastructure (EDI DevOps):** platform for requirement management, code development, testing and deployment to operations
- **User Support Tool (UST):** user relationship management and issue resolution (Kayako)
- **Earthdata Log-in (User Registration System):** provides a centralized and simplified mechanism for user registration and account management for all EOSDIS system components.



Common Metadata Repository

Provides a single source of unified, high-quality, and reliable Earth Science metadata with a high performance ingest and search architecture for submission and discovery of all EOSDIS data sets.



Lightning fast, always available

- 95% queries complete in <1s
- 99.98% uptime (last 365d)

Big Data Ready

- 34K collections
- 367 million files indexed
- Prepared to scale 1B+ records

Standards-focused

- ISO-19115 metadata
- OpenSearch/OGC CSW
- REST based APIs

Community-focused

- Developer's portal
- Active Developer's forum
- Ecosystem of supported tools
- Open Source codebase in github

Internationally Recognized

- Provides the backbone of the Community of Earth Observing Satellites International Directory Network (CEOS IDN)

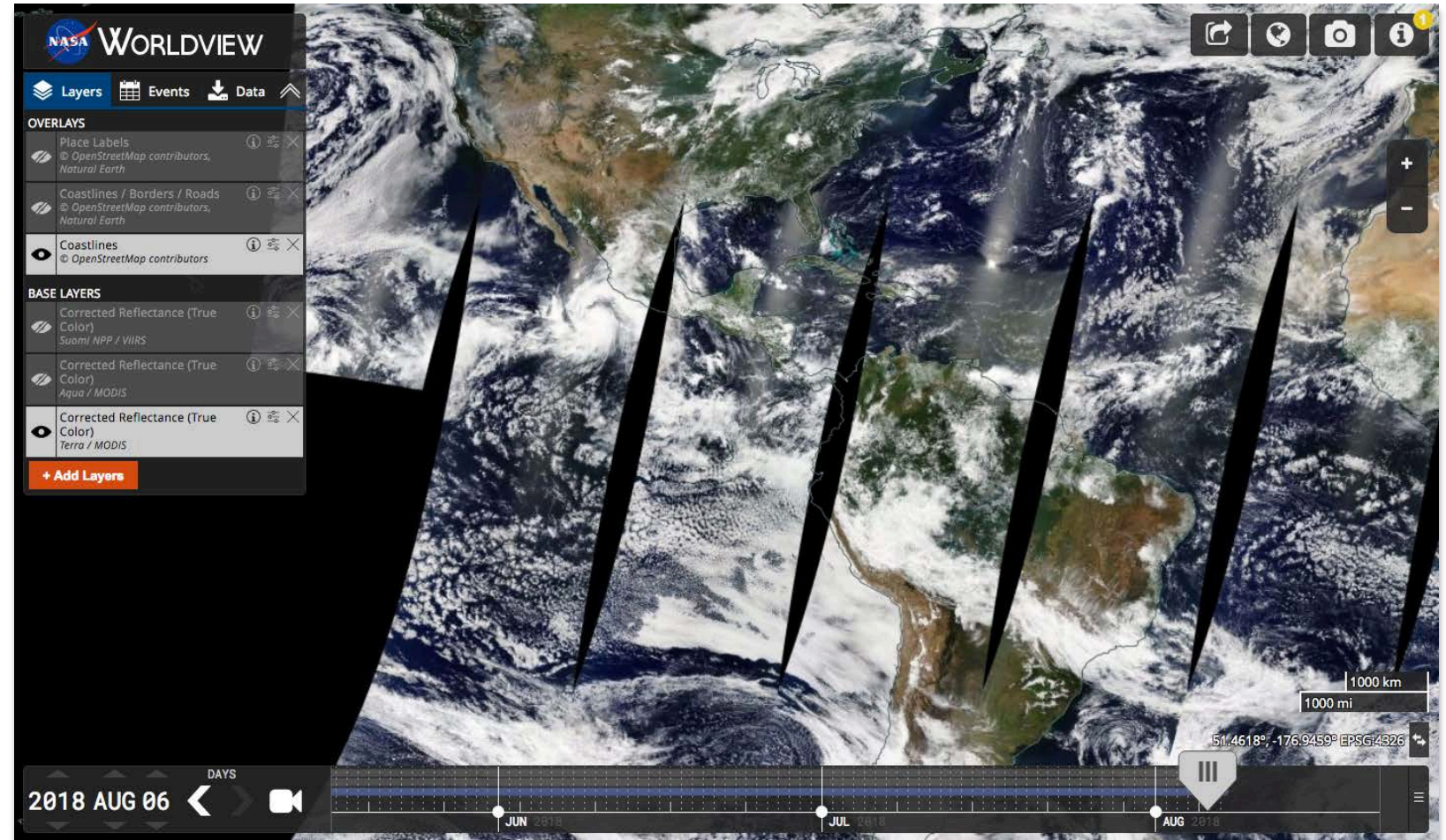
Data Discovery and Access

- Earthdata Search facilitates discovery of data across all 12 DAACs (search.earthdata.nasa.gov)
- Data is online and can be readily downloaded

The screenshot displays the Earthdata Search web application. On the left, a sidebar lists various instruments and platforms, with 'AIRS' and 'ASTER' checked. The main area features a search bar with the text 'Type any topic, collection, or place name'. Below the search bar, a date range is set from '2006-01-01 00:00:00' to '2008-12-31 23:59:59', and a rectangular geographic area is defined with coordinates: SW: 26.015625, -97.3125 and NE: 34.3125, -86.90625. A map of North America shows a red rectangle over the central United States. Below the map, it indicates '71 Matching Collections'. Two collection entries are visible: 'ASTER Level 1 precision terrain corrected registered at-sensor radiance V003' and 'AIRS/Aqua L3 Daily Standard Physical Retrieval (AIRS+AMSU) 1 degree x 1 degree V006 (AIR3STD) at GES DISC'. The footer includes version information (v 1.53.5), search time (4.9s), and NASA contact information.

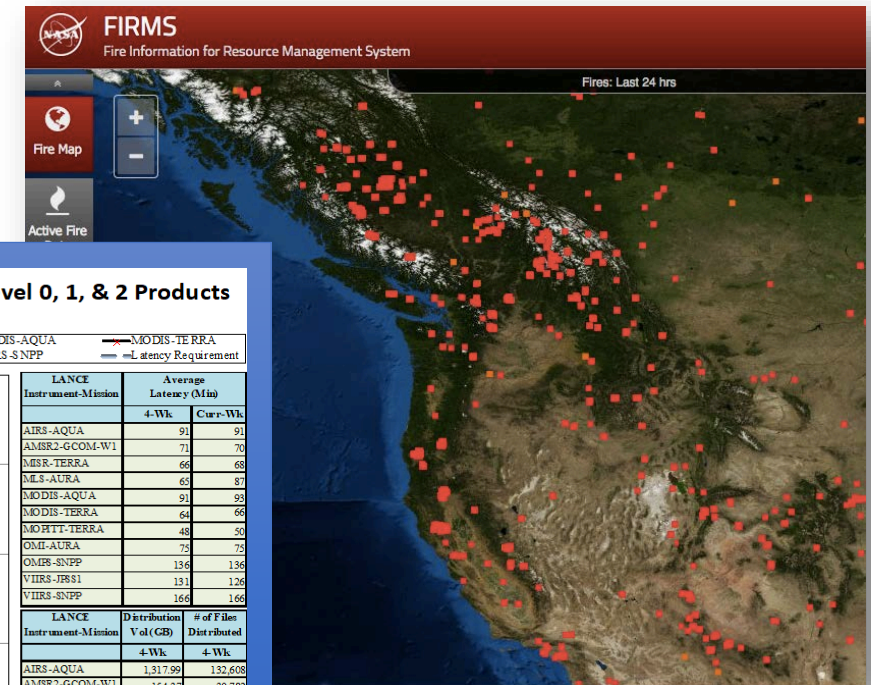
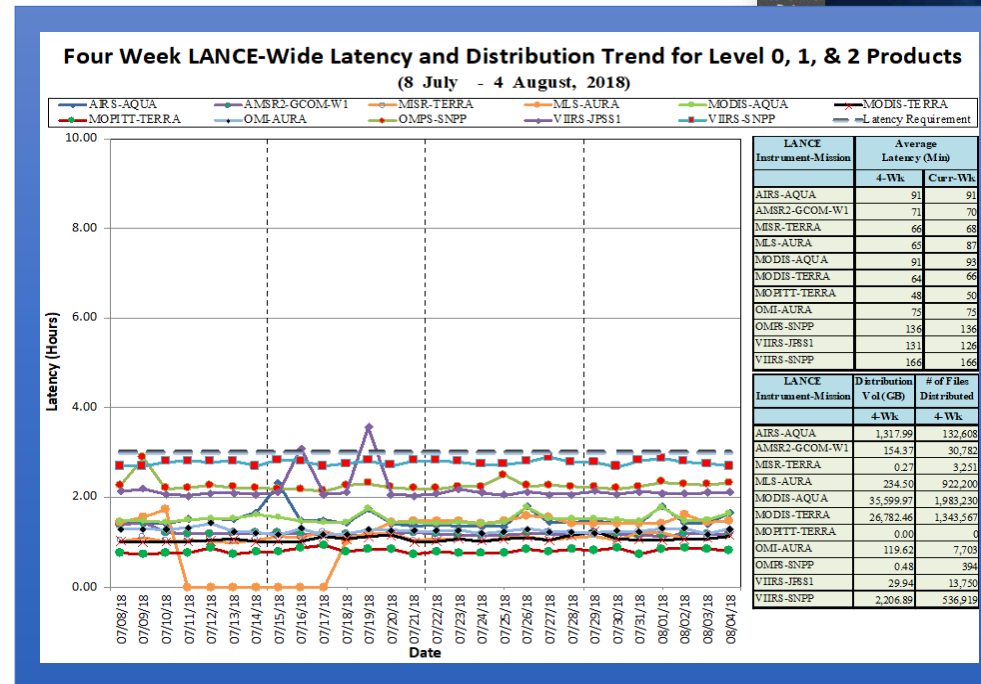
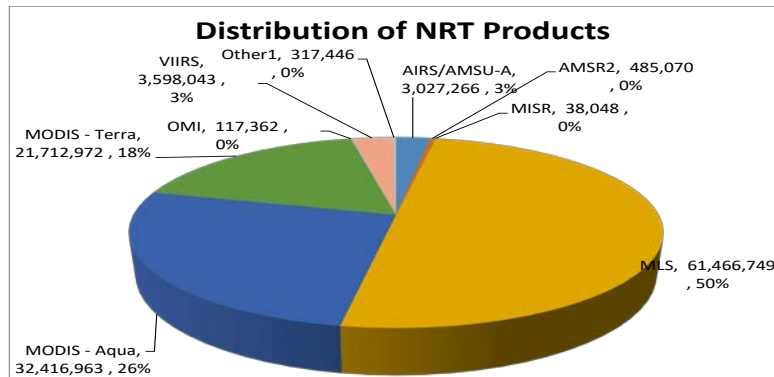
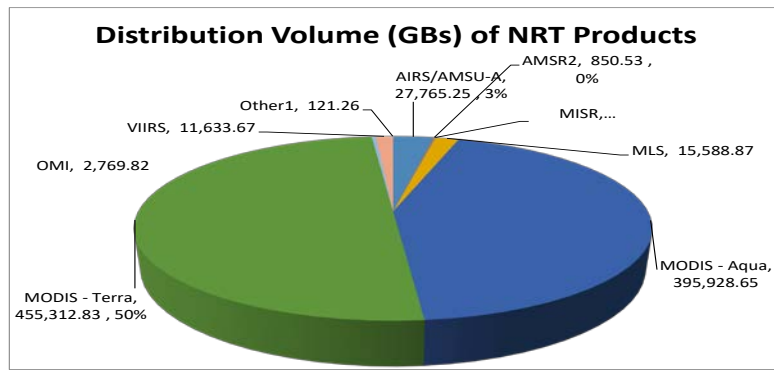
Full-Resolution Browse Capabilities

- Global Imagery Browse System (GIBS) aids in discovery and access
- GIBS software is open source; anyone can develop clients; Worldview is NASA's client



Near-Real Time Data Support

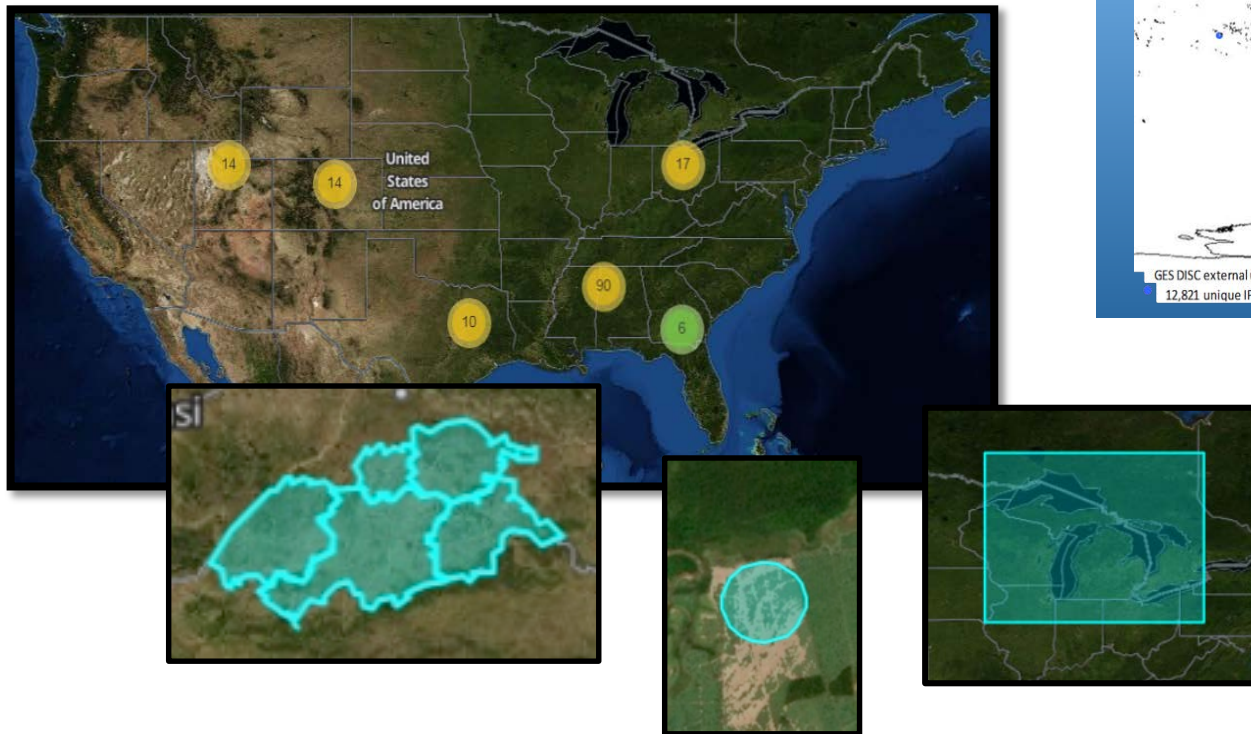
- More than 380 unique datasets available within 3 hours of observation to serve a growing applications community
- In FY17, EOSDIS distributed over 1 Petabyte of data (123+ million files) to over 370,000 users



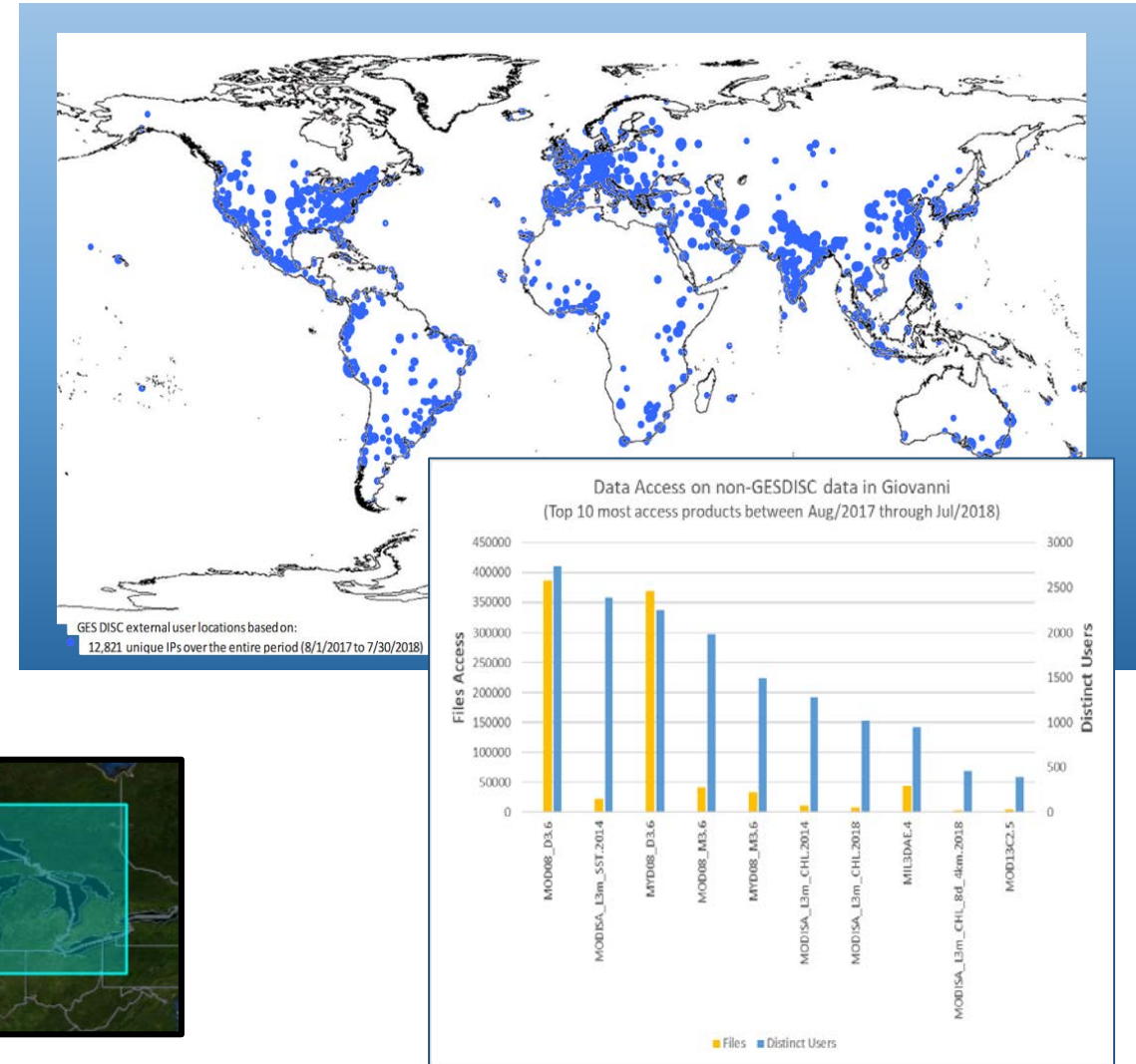
Data Analysis Tools

Users can discover, analyze and visualize hundreds of products using technique customized for their science discipline

Land Processes DAAC (USGS) - AppEEARS

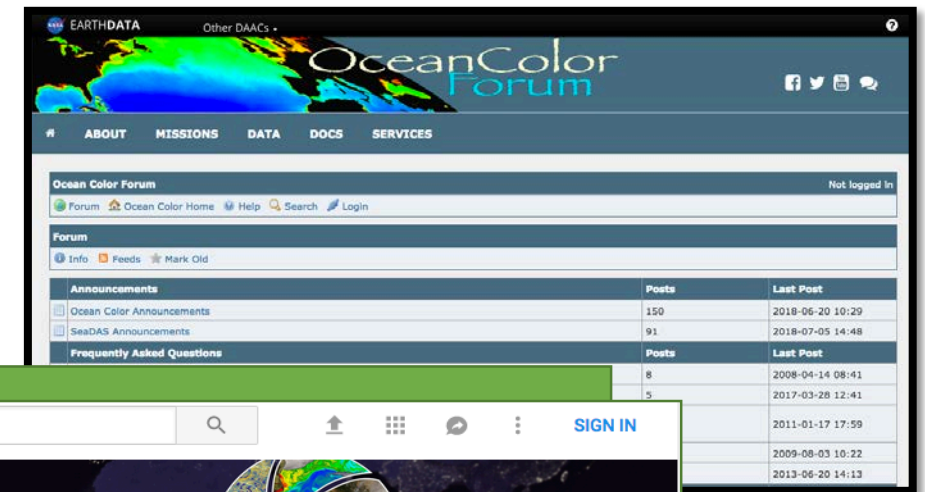


Goddard DAAC (GES DISC) - GIOVANNI



User Support and Services

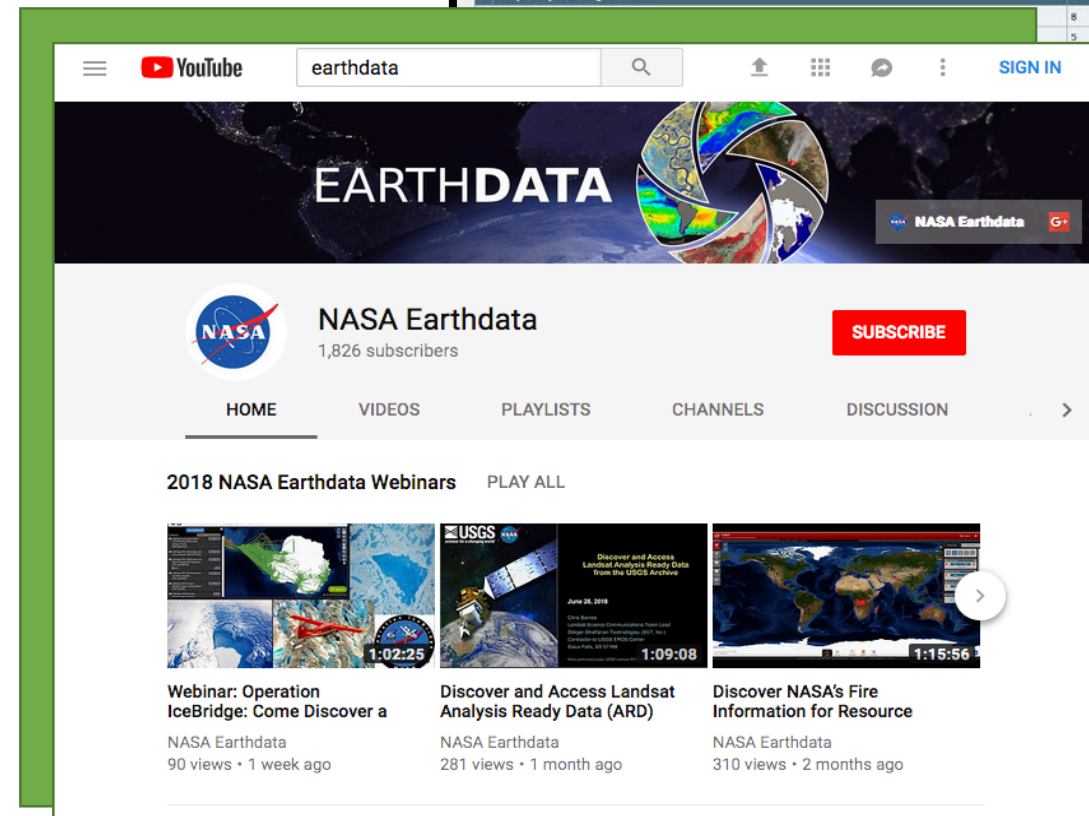
- Each DAAC has a user services group to address users' questions about data
- Frequently Asked Questions (FAQ) Pages
- User Forums for information exchange
- Webinars and tutorials
- Data recipes
- Specific documentation



The screenshot shows the Ocean Color Forum website. At the top, there is a navigation menu with links for ABOUT, MISSIONS, DATA, DOCS, and SERVICES. Below the menu, there is a search bar and a login link. The main content area features a table with two sections: 'Announcements' and 'Frequently Asked Questions'. Each section has columns for 'Posts' and 'Last Post'.

Announcements	Posts	Last Post
Ocean Color Announcements	150	2018-06-20 10:29
SeaDAS Announcements	91	2018-07-05 14:48

Frequently Asked Questions	Posts	Last Post
	8	2008-04-14 08:41
	5	2017-03-28 12:41
		2011-01-17 17:59
		2009-08-03 10:22
		2013-06-20 14:13



The screenshot shows the NASA Earthdata YouTube channel page. The channel name is 'NASA Earthdata' with 1,826 subscribers. The page features a navigation menu with links for HOME, VIDEOS, PLAYLISTS, CHANNELS, and DISCUSSION. Below the menu, there is a section titled '2018 NASA Earthdata Webinars' with a 'PLAY ALL' button. Three webinar videos are displayed, each with a thumbnail, title, and view count.

Webinar Title	View Count	Time Ago
Webinar: Operation IceBridge: Come Discover a	90 views	1 week ago
Discover and Access Landsat Analysis Ready Data (ARD)	281 views	1 month ago
Discover NASA's Fire Information for Resource	310 views	2 months ago

NASA's Earth Science Data System in 2018



EOSDIS currently has over **27 Petabytes** of accessible Earth science data

Easy access and discovery of data to over **12,500 unique data products**

... of which 95% of granule searches complete in less than **1 Second**



EOSDIS delivered over **1.6 Billion** data products to over **4.1 Million** users from around the world

33,000 Data Collections in the Common Metadata Repository (CMR)



EOSDIS also delivers near-real-time products in under **3 hours** from observation ...

Over **330,000 users** have registered with EOSDIS to date



And Over **380 Million** data granules



American Customer Satisfaction Index (ACSI) survey scoring **79** from over **4,000** respondents



Data Discovery

Metadata

User Growth

New, easy to use software, tools, services and data formats have exposed EO data to an ever growing user base.

Includes 2 user types:

Local Users

- Very knowledgeable about the specific scientific context within which data were collected
- Don't require as much contextual information to find and use relevant data
- Includes:
 - Domain specific research scientists
 - Principal investigators who originally collected the data

Global Users

- Leverage data for research and applications beyond the data's original intended use
- Includes:
 - Scientists conducting research across siloed domain environments
 - Users from the applications and decision making communities
 - Data scientists using data in innovative new ways

Where Do Data and Users Come Together?

- For local users -> local data centers
- For global users -> Centralized, or aggregated catalogs
 - Provides a single discovery point for data from multiple sources
 - Brings together metadata from different data centers into an aggregated catalog and presents the metadata in a unified user interface
- NASA's aggregated catalog for Earth observation data is the Common Metadata Repository (CMR) and the unified user interface is the Earthdata Search client.



Metadata In Aggregated Catalogs

Metadata sets the stage for data -

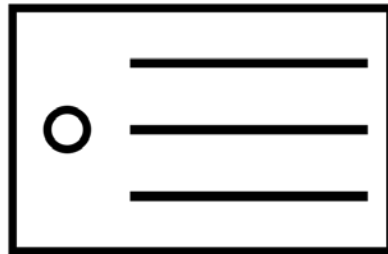
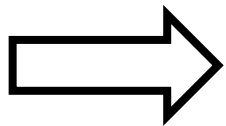
- Metadata makes it possible to search for data
- Metadata limits and focuses attention to the relevant information about a dataset
- Metadata helps a user understand whether data is relevant to a given research problem

When metadata isn't at its best, users can't –

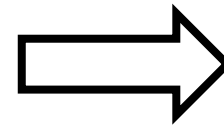
- Find the right data
- Understand the data



Created by ProSymbols
from Noun Project



Created by Guilhem
from Noun Project



Created by Marksu Desu
from Noun Project

1. Tag by Guilhem from the Noun Project
2. Big data by ProSymbols from the Noun Project
3. users by Marksu Desu from the Noun Project

When Metadata Doesn't Work...

- Conducting a faceted search for 'NDVI' in Earthdata Search returns 14 datasets
- NDVI, or the Normalized Difference Vegetation Index, is key for many applications based research questions
- MODIS datasets are missing from the search results:
- MODIS is a key instrument for calculating NDVI
- MODIS Level 3 vegetation indices datasets which include NDVI as a vegetation layer are missing

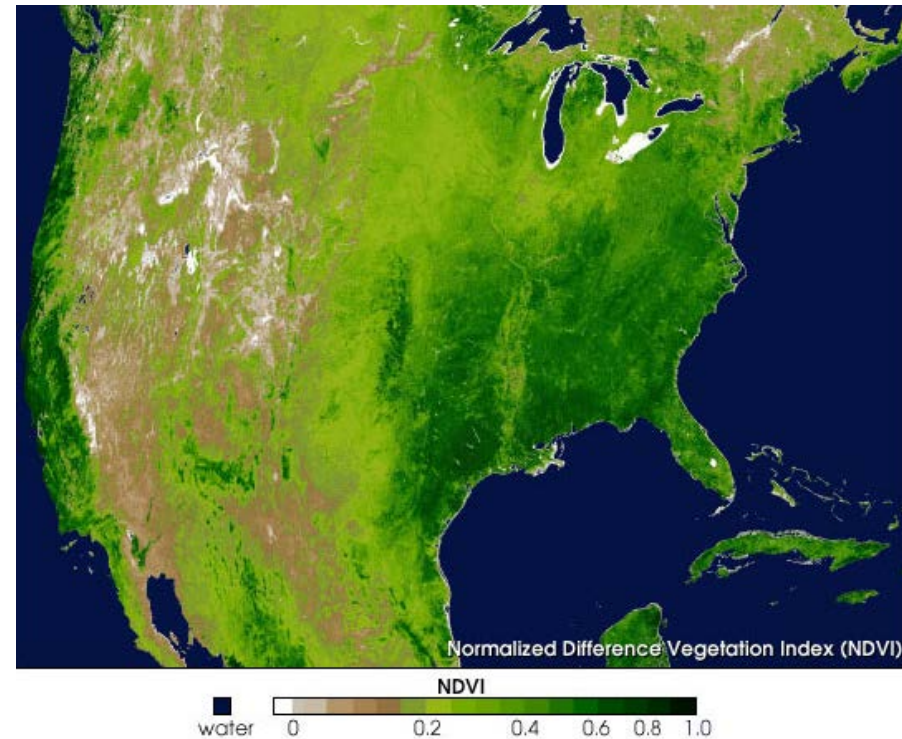
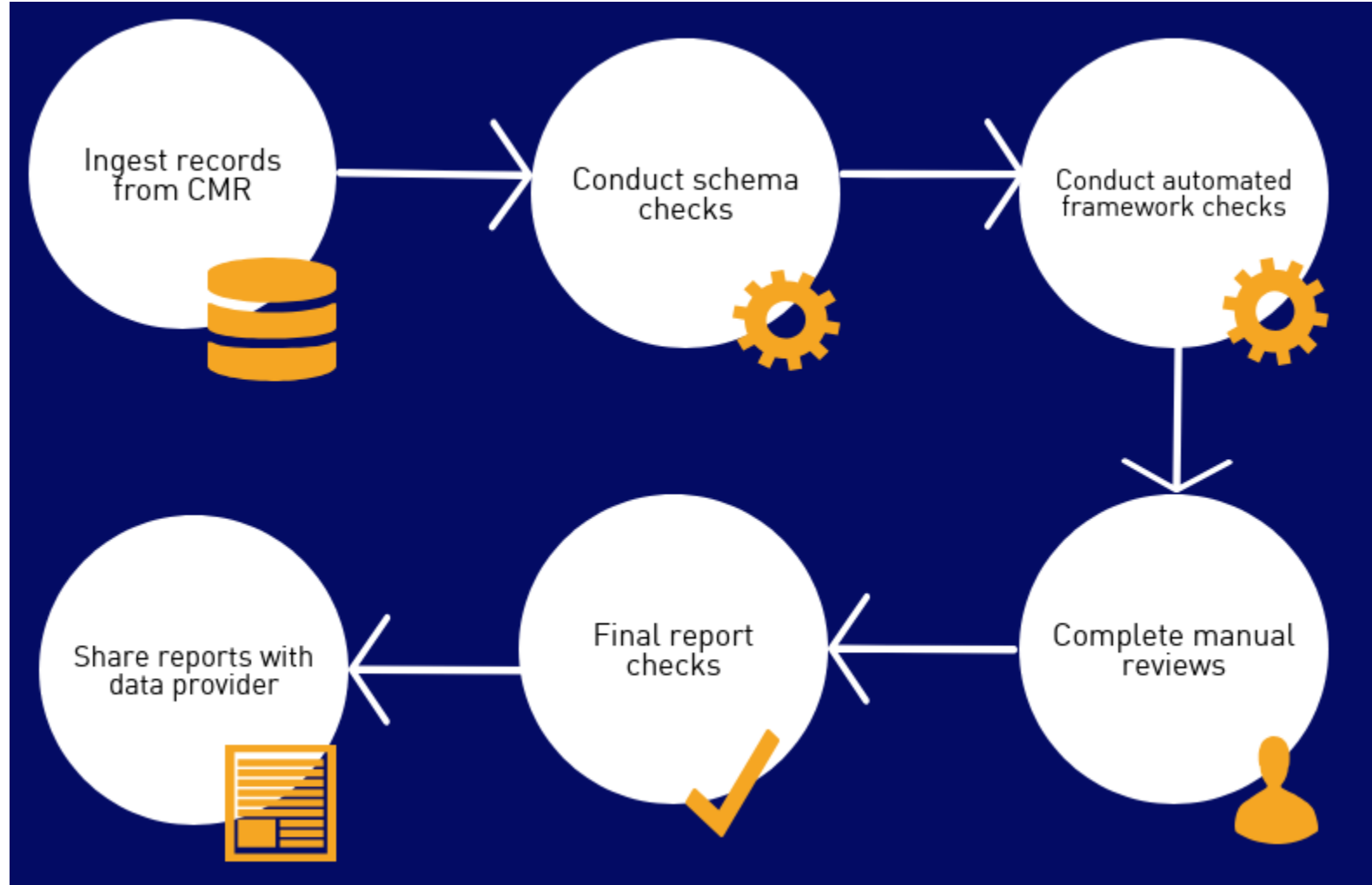


Image Credit: <https://earthobservatory.nasa.gov/images/696/spring-vegetation-in-north-america>

How Can We Assess Metadata Quality?

- Metadata needs to be informative to both subject matter experts and applications based, or global, users
- High quality metadata helps both user groups
- However, creating and maintaining high quality metadata can cause metadata friction for data centers
- NASA has established the Analysis and Review of CMR (ARC) team to define and assess metadata quality for Earth observation data and to lower metadata friction for data centers by:
 - Creating a metadata quality framework to assess metadata quality consistently and rigorously
 - Leveraging automated and manual checks to assess this quality
 - Building a team of reviewers with backgrounds in Earth system science, Atmospheric science, remote sensing and informatics
 - Defining a priority matrix to help lower metadata friction for data centers

ARC Metadata Quality Review Process



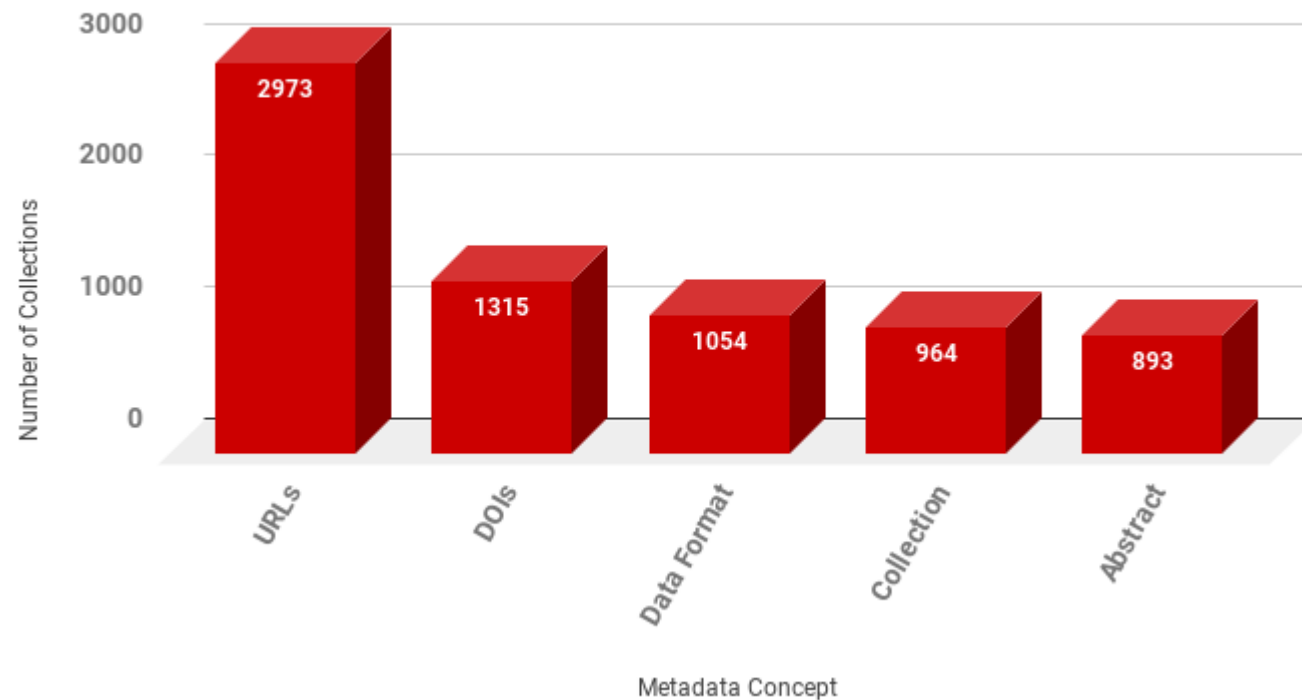
- Leverages metadata curation framework and ARC priority matrix
- Process followed for each collection level metadata record and one randomly selected granule/file level record

ARC Priority Matrix

Priority Categorization	Justification
Red = High Priority Issues	High priority issues emphasize several characteristics of metadata quality including completeness, accuracy and accessibility. Issues flagged as red are required to be addressed by the data provider.
Yellow = Medium Priority Issues	Medium priority issues emphasize consistency and completeness. Data providers are strongly encouraged to address yellow flagged issues. If a yellow flagged issue is not addressed, the data provider will be asked to provide a justification as to why.
Blue = Low Priority Issues	Low priority issues also focus on completeness, consistency and accuracy. Any additional information that may be provided to make the metadata more robust or complete is categorized as blue.
Green = No Issue	Elements flagged green are free of issues. Green flagged elements require no action on behalf of the data provider.

Top Metadata Quality Issues

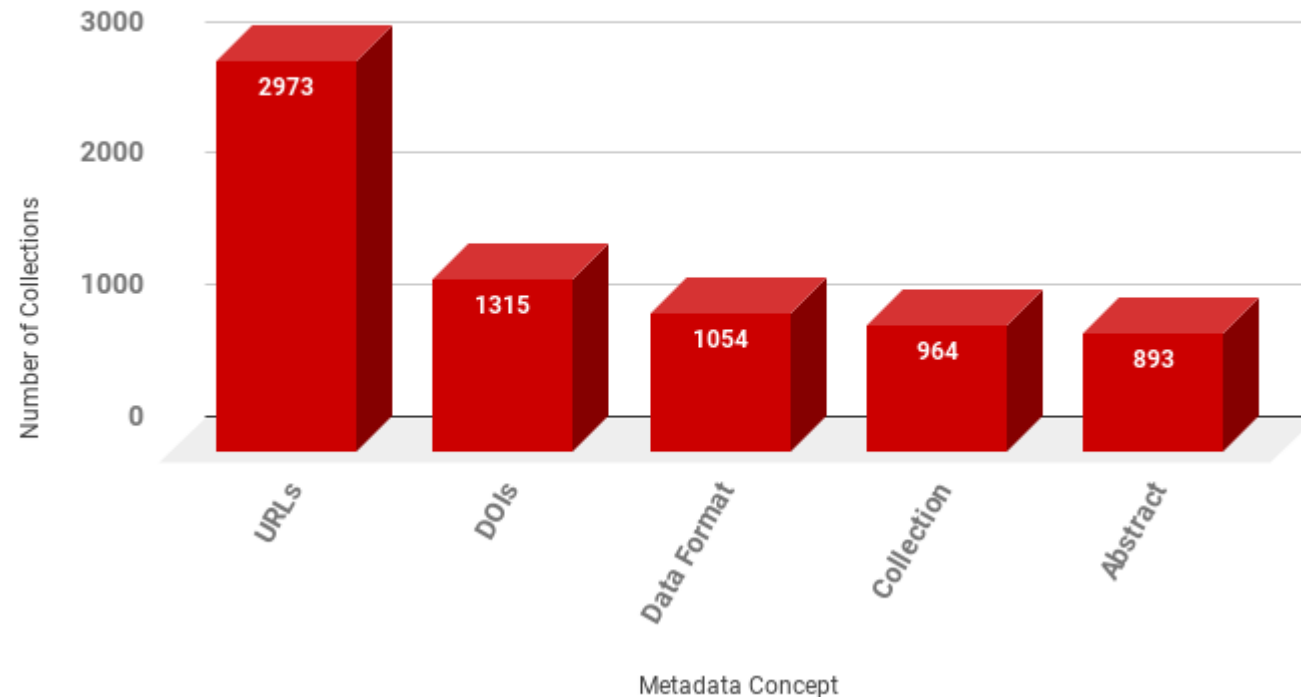
Top 5 Red Errors (1929 Records)



- Broken URLs:
- Data access URLs that do not conform to NASA requirements (ftp vs https)
- No URLs to essential data documentation
- *No data access URLs provided at all*

Top Metadata Issues

Top 5 Red Errors (1929 Records)

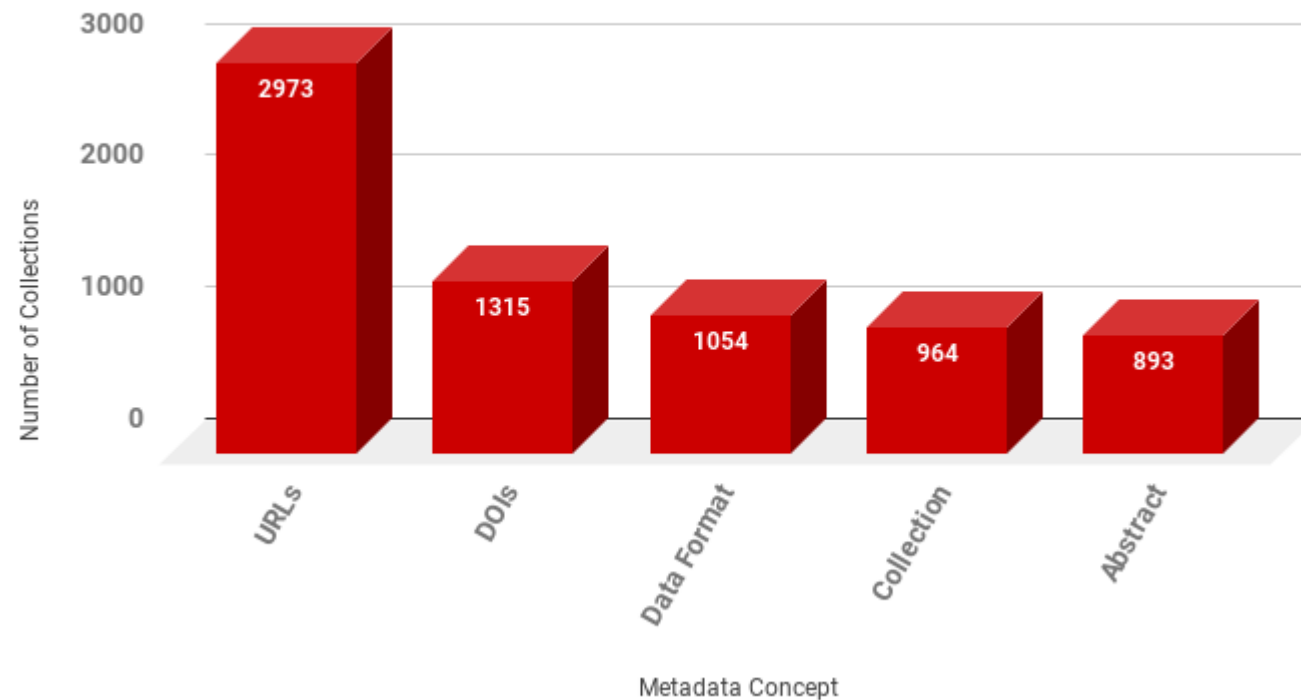


- DOIs and Collection State are new concepts that were recently added
- Slow adoption of new concepts by data centers despite creating DOIs for data

Top Metadata Issues



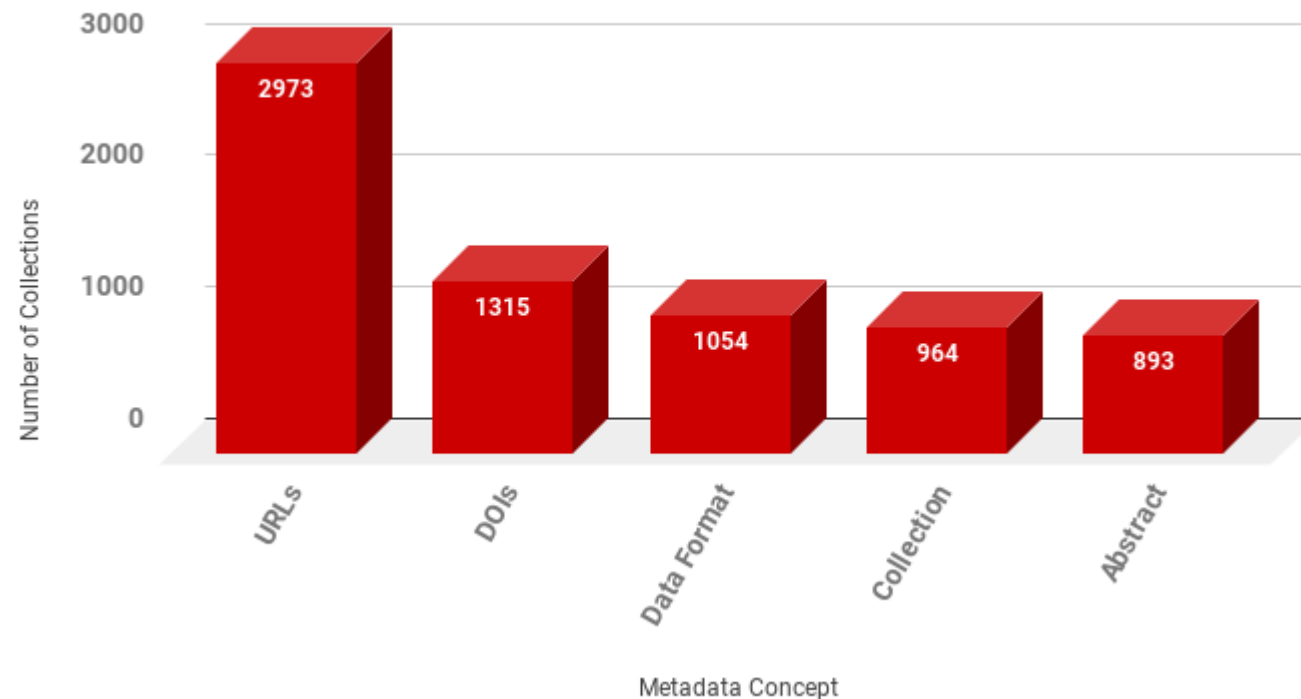
Top 5 Red Errors (1929 Records)



- Data format information not widely adopted by data centers
- Not viewed as an information priority in the past but important to users

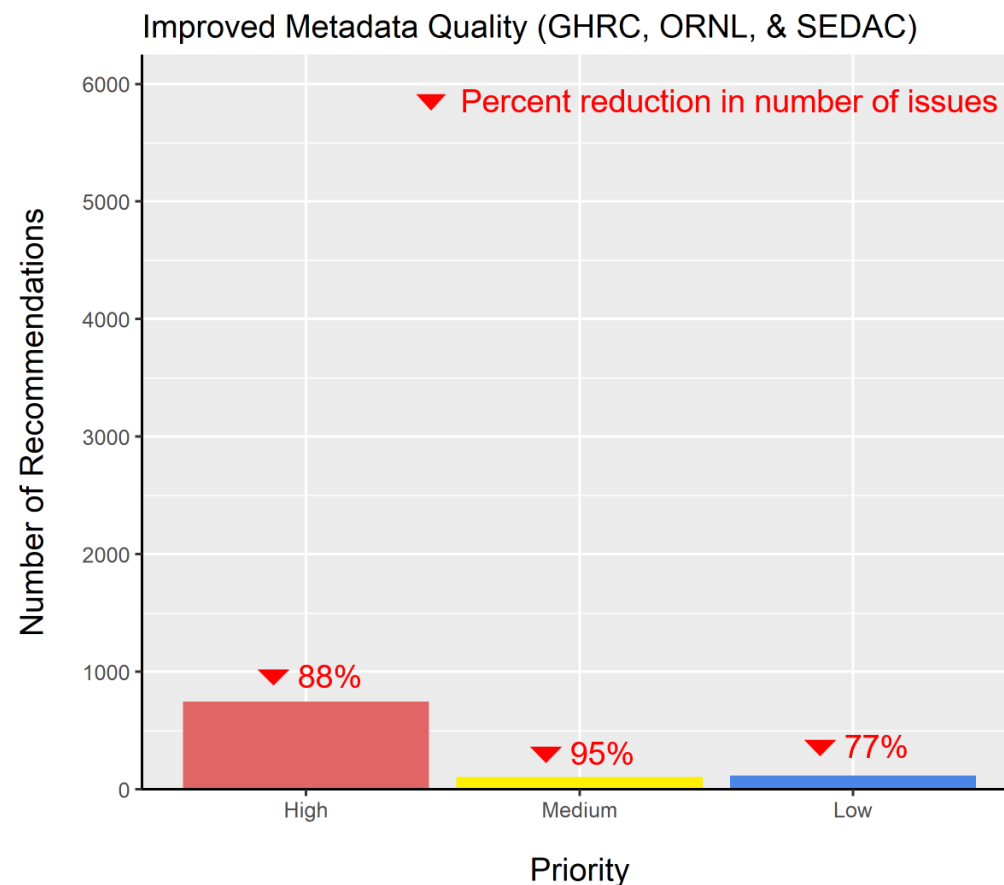
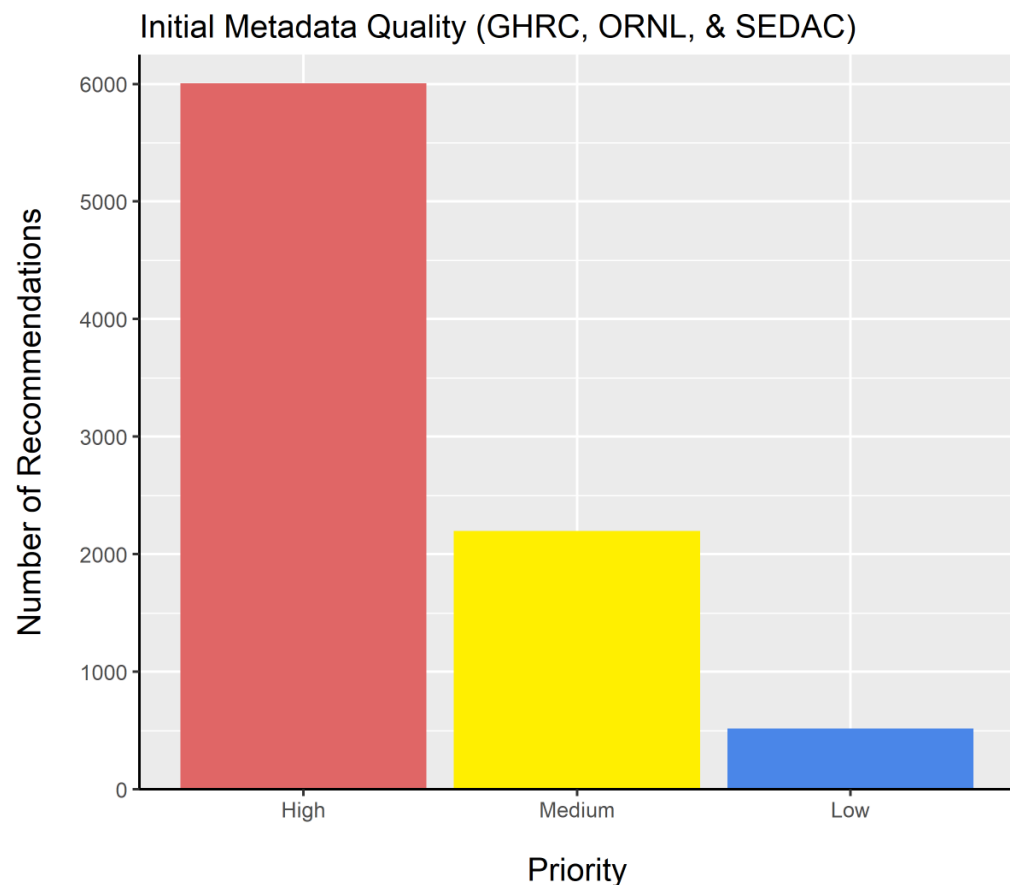
Top Metadata Issues

Top 5 Red Errors (1929 Records)



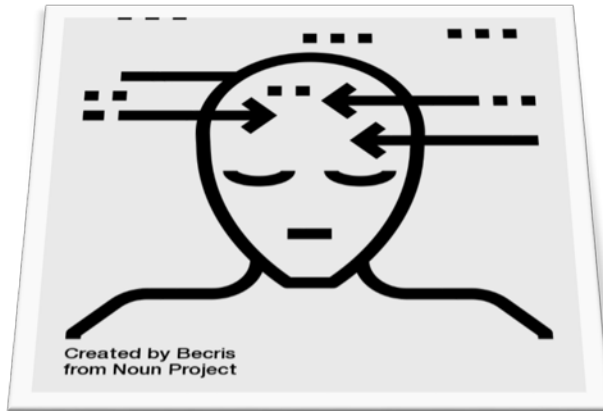
- Abstracts are particularly problematic
 - Too lengthy
 - Non-existent
 - Not specific enough to describe data
 - Too technical for a global user

Metadata Improvement To Date



Sample of improved metadata quality to date (3 of 12 data centers)
All data centers are actively participating in the improvement effort

Lessons Learned



Learning by Becris from the Noun Project

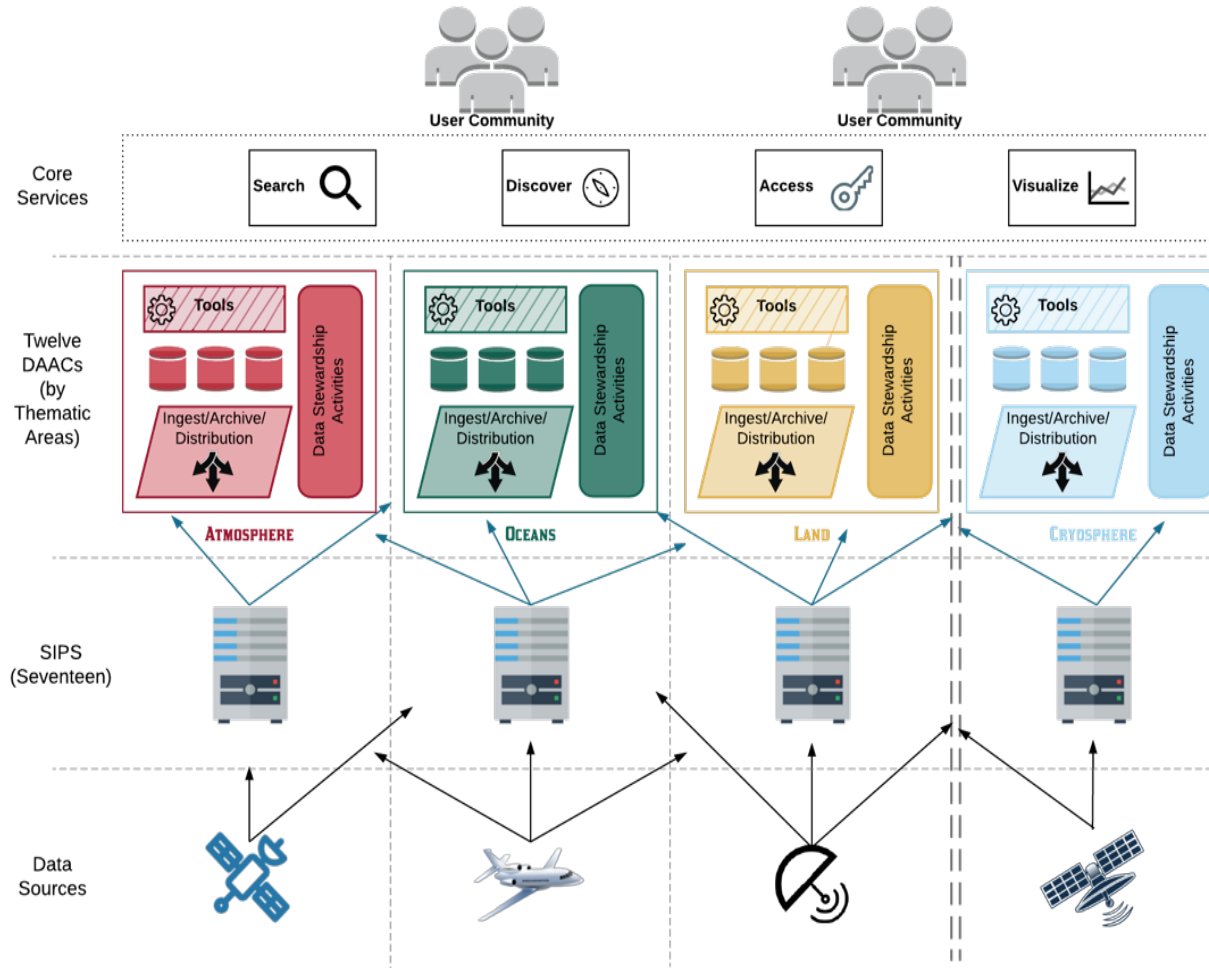
1. Leveraging a metadata quality framework operationally requires communication, compromise and reiteration
2. The metadata curation process is not a “Do-it-once-right-and-forget-about-it” activity but should instead be viewed as an iterative process
3. Curating metadata within an aggregated catalog may require an organizational mindset change
 - Need to consider not just local users needs and local metadata needs when curating metadata



Data Use

*Cloud Infrastructure
Analytics*

Current EOSDIS Architecture



Discipline specific support and tools (**DAAC data**)
Optimized for archive, search and distribution
Easily add new data products
Supports millions of users – High ACSI score

Uneven service and performance
Significant interface coordination
Limited on-demand product generation
Fragmentation – duplication of services, software and storage

Open data policies drive system architecture

NASA Earth Science Missions: Present through 2023

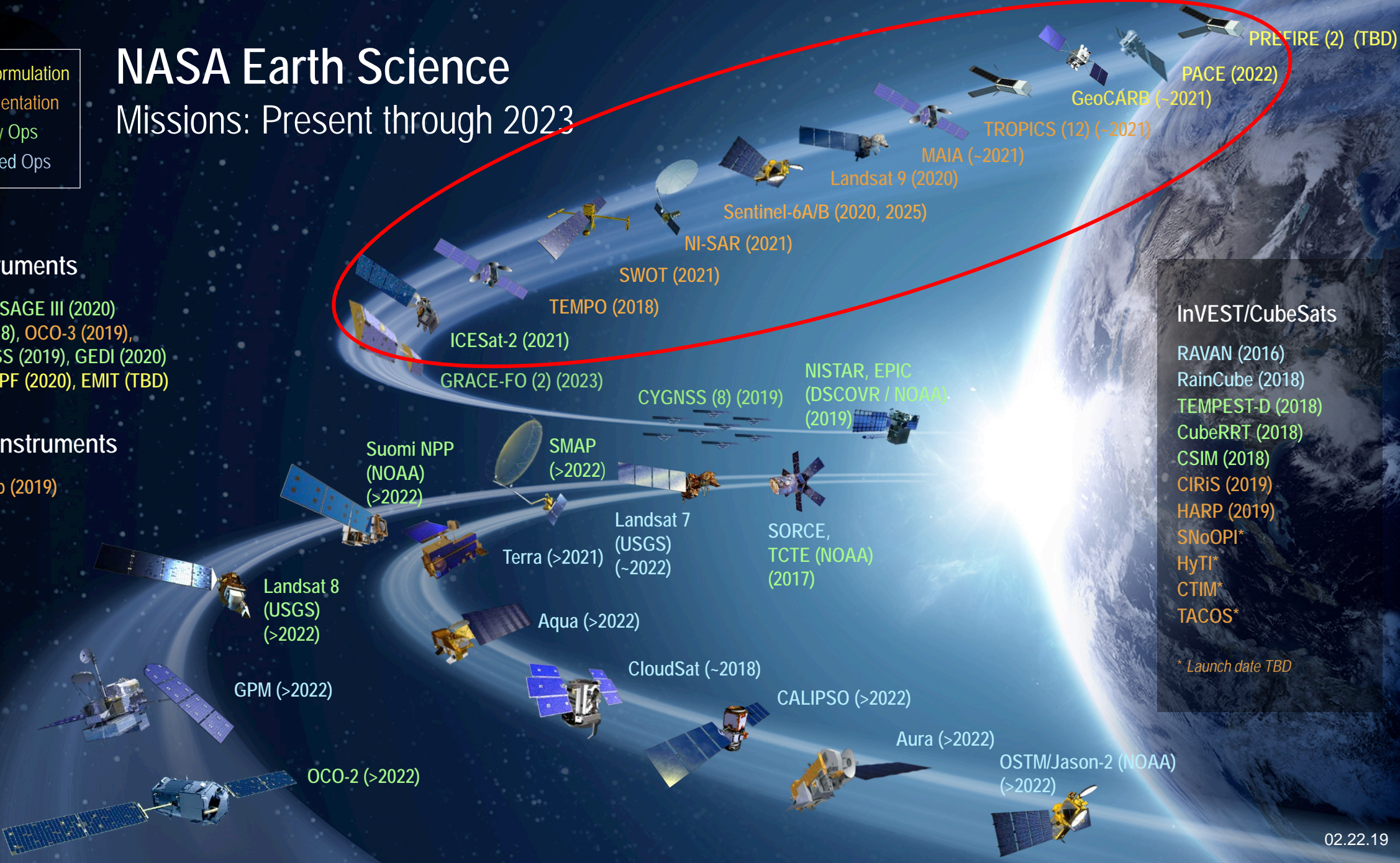
- (Pre)Formulation
- Implementation
- Primary Ops
- Extended Ops

ISS Instruments

LIS (2020), SAGE III (2020)
 TSIS-1 (2018), OCO-3 (2019),
 ECOSTRESS (2019), GEDI (2020)
 CLARREO-PF (2020), EMIT (TBD)

JPSS-2 Instruments

OMPS-Limb (2019)



InVEST/CubeSats

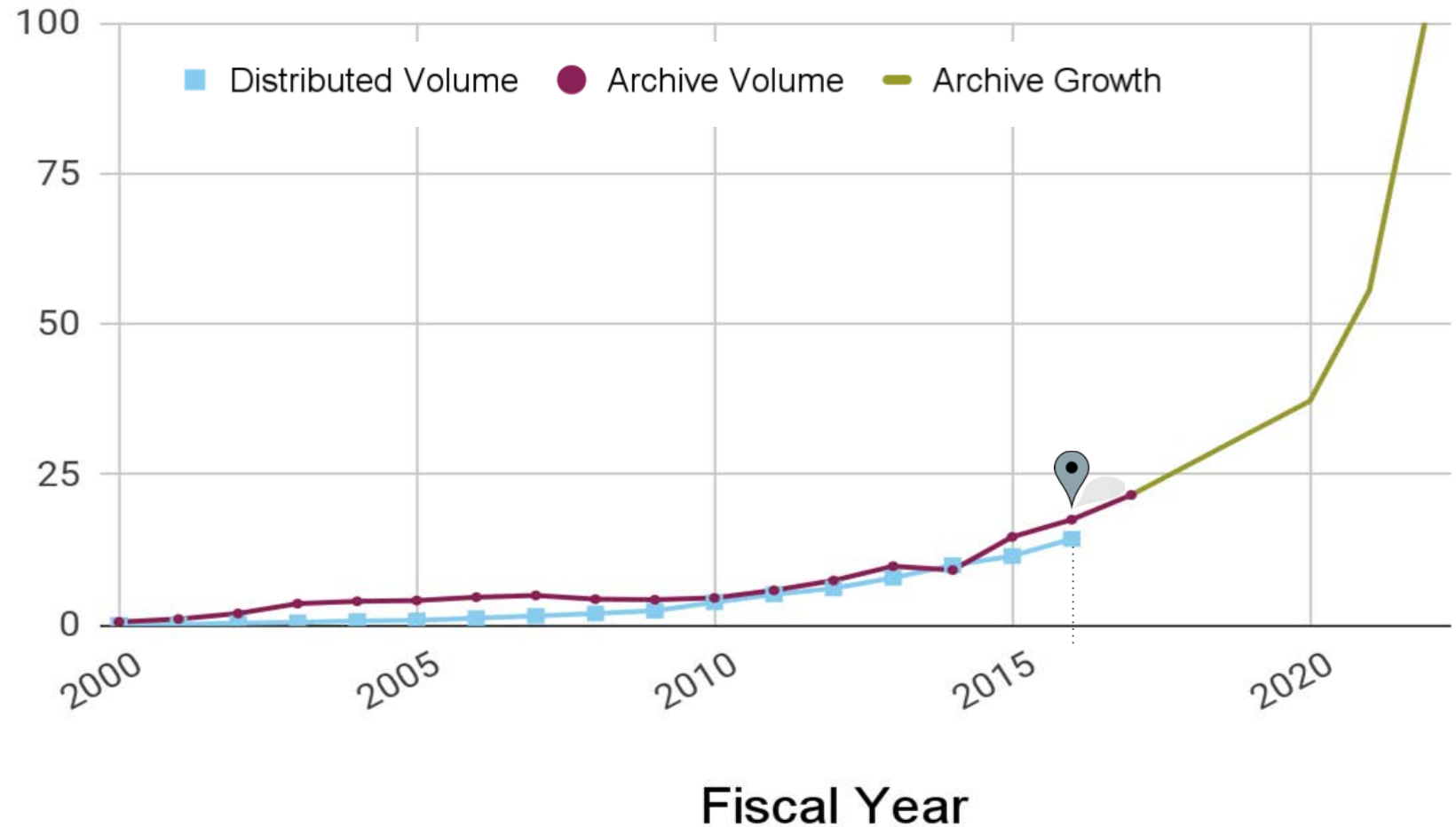
- RAVAN (2016)
- RainCube (2018)
- TEMPEST-D (2018)
- CubeRRR (2018)
- CSIM (2018)
- CIRiS (2019)
- HARP (2019)
- SNoOPI*
- HyTI*
- CTIM*
- TACOS*

** Launch date TBD*

EOSDIS Data System Evolution

The current architecture will not be cost effective as the annual ingest rate increases from 4 to 50PB/year

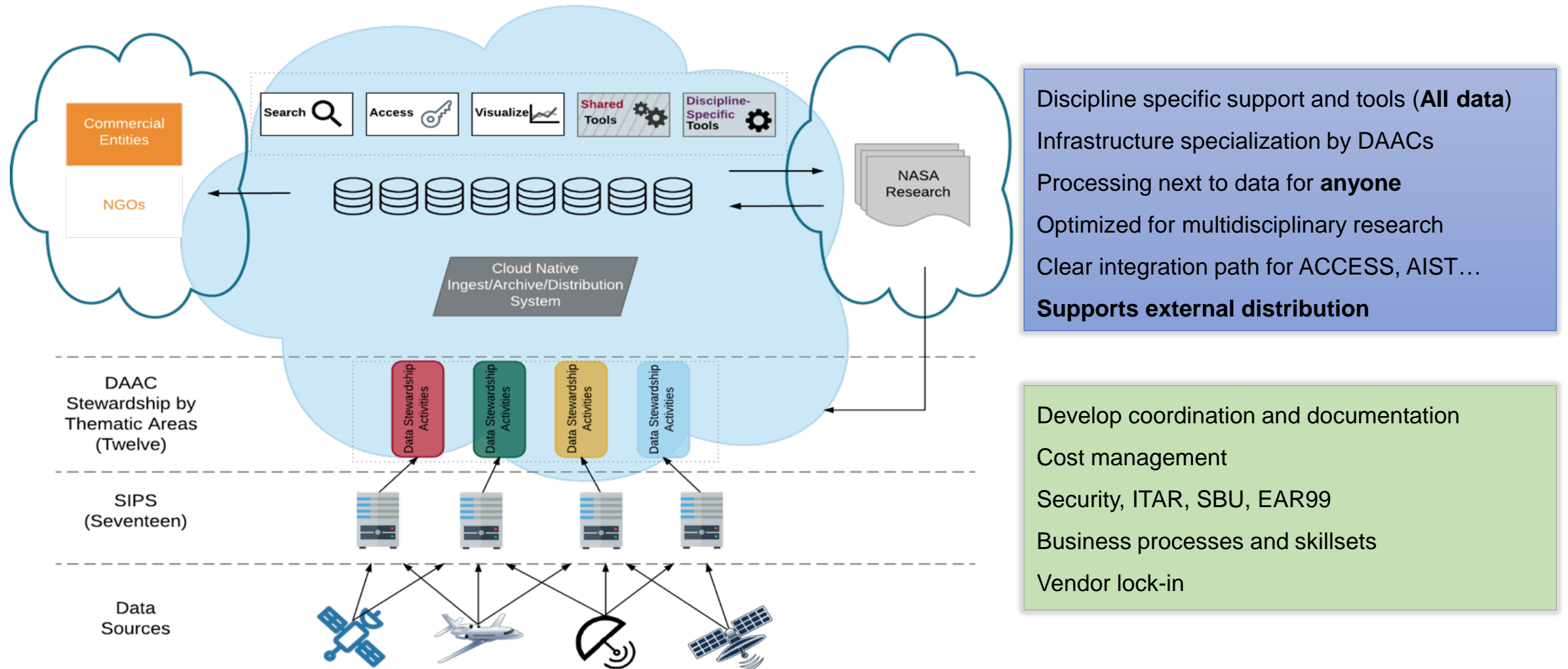
EOSDIS is developing open source cloud native software for reuse across the agency, throughout the government and for any other user.



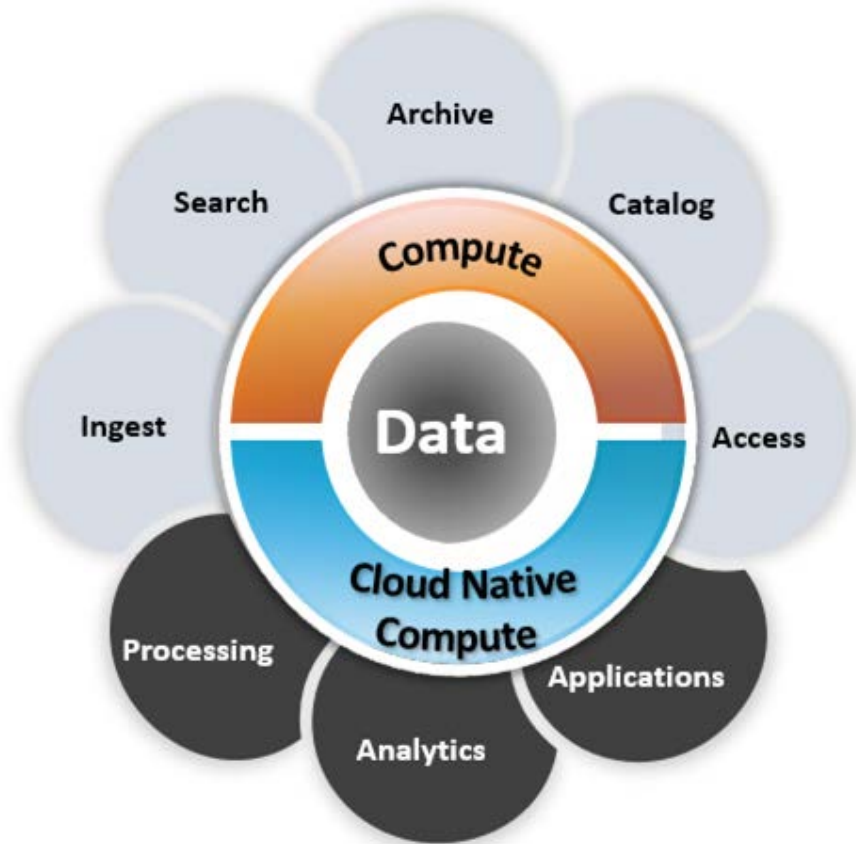
Cloud offers benefits like the ability to analyze data at scale, analyze multiple data sets together easily and avoid lengthy expensive moves of large data sets allowing scientists to work on data "in place"

Simplified EOSDIS Cloud Architecture - 2021

Open Science = Open Data + Open Source Software + Open Services



Conceptual 'data close to compute'



The operational model of consolidating data—allowing users to compute on the data in place with a platform of common tools—is natural to cloud; it is a cost-effective way to leverage cloud and could be applicable to many businesses and missions

Bring customers to the data

Large volume data storage: Centralized mission observation and model datasets stored in auto graduated AWS object storage (Amazon S3, Amazon S3 IA, Amazon Glacier)

Scalable compute: Provision, access, and terminate dynamically based on need. Cost by use

Cloud Native Compute: Cloud vendor service software stacks and microservices easing deployment of user based applications

EOSDIS applications and services: Application and service layer using AWS compute, storage (Amazon S3, Amazon S3 IA, Amazon Glacier), and cloud native technologies

Non-EOSDIS/public applications and services: Science community brings algorithms to the data. Support for NASA and non-NASA

So we made this thing.



CUMULUS

What is Cumulus?

Lightweight, cloud-native framework for data ingest, archive, distribution and management

Goals

- Provide core DAAC functionality in a configurable manner
- Enable DAACs to help each other with re-usable, compatible containers (e.g. data retrieval, metadata extraction, metrics delivery)
- Enable DAAC-specific customizations

Cumulus Major System Components

A lightweight framework consisting of:

Tasks a discrete action in a workflow, invoked as a Lambda function or EC2 service, common protocol supports chaining

Orchestration engine (AWS Step Functions) that controls invocation of tasks in a workflow

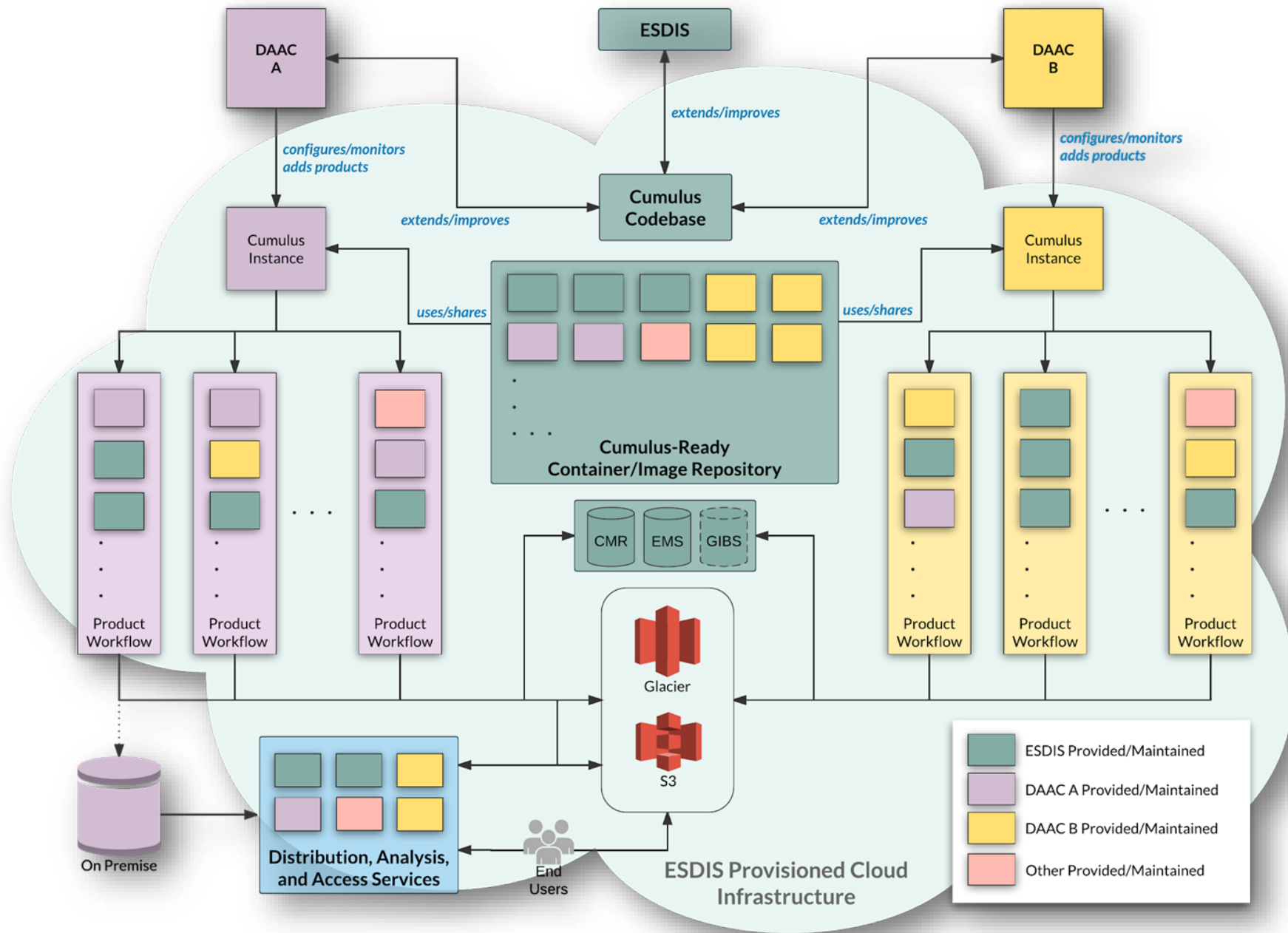
Database store status, logs, and other system state information

Workflows(s) file(s) that define the ingest, processing, publication, and archive operations (json)

Dashboard create and execute workflows, monitor system

The screenshot displays the Cumulus dashboard interface. At the top right, a workflow diagram shows a sequence of steps: 'Start' (yellow circle), 'Ingest' (green rectangle), 'Process' (green rectangle), 'Publish' (green rectangle), 'Archive' (green rectangle), 'End' (yellow circle), and 'Update' (green rectangle). Below the diagram, the dashboard features a 'Dashboard' section with several metrics: 0 Errors, 7 Collections, 0 Granules Processed in the Past Hour, 51.79s Average Processing Time, 0 Pending Tasks, 14 Running Tasks, and 4 Queued Messages. A 'Collections' section shows 'Granules Updated Today (121K)' with 0 Granules Ingesting, 45 Granules Processing, 0 Granules Updating CMR, and 0 Granules Archiving. Below this, a table lists various collections with columns for Name, Granules, Completed, Failed, Average Duration, Created at, and Updated at.

Name	Granules	Completed	Failed	Average Duration	Created at	Updated at
MYD13A1_version_006	2,627	2,612	0	-12.84s	16:52:11 04/04/17	11:00:29 06/27/17
MYD09A1_version_006	5,076	5,044	0	-7.32s	17:38:29 04/04/17	11:00:29 06/27/17
MOD14A1_version_006	4,710	4,701	0	-3.11s	16:43:20 04/04/17	11:00:29 06/27/17
MOD11A1_version_006	1	1	0	-109.03s	19:56:53 11/25/73	11:00:29 06/27/17
MOD09GQ_version_006	40,960	40,556	234	25.55s	17:25:11 04/04/17	11:00:29 06/27/17
AST_L1A_version_003	49,033	49,022	11	11.88s	19:56:53 11/25/73	11:00:29 06/27/17

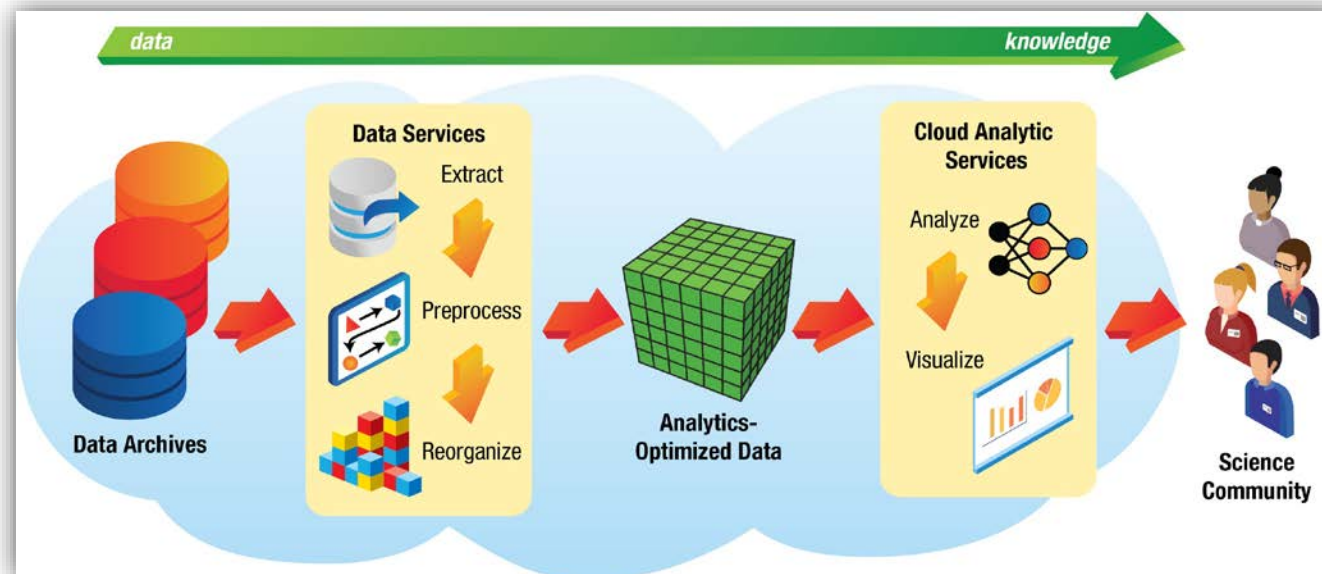


Workshop- “Enabling Analytics in the Cloud for Earth Science Data”

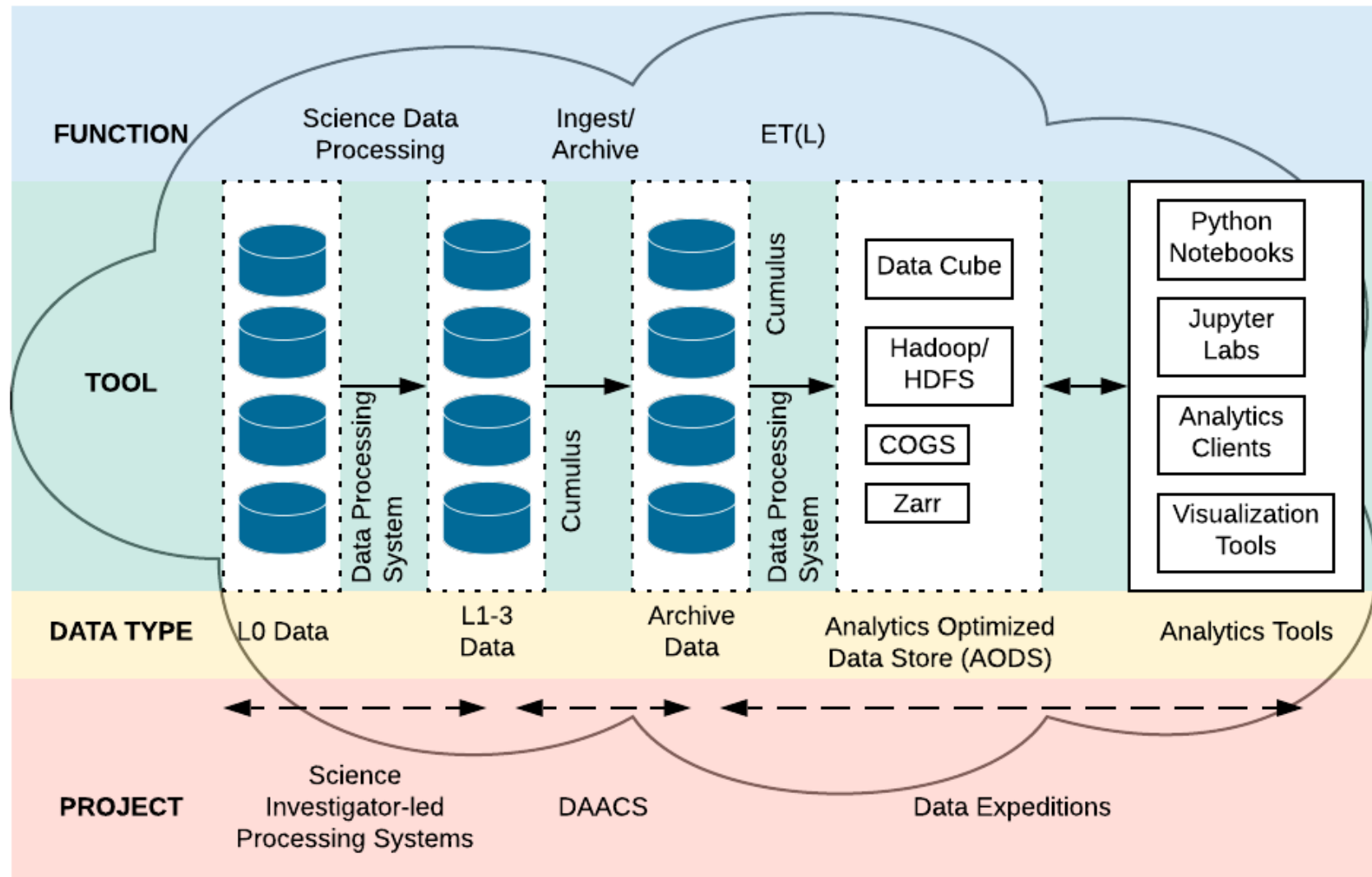
- The NASA Earth Science Data Systems Program sponsored a workshop on 21-23 February 2018 where the participants discussed the convergence of Big Data, Cloud Computing, and Analytics.
- These discussions clustered around these themes
 - Strategic Alignment of Cloud Effort
 - Reference Architecture for Cloud Analytics
 - **Analytics Optimized Data Stores**
 - Reuse of Cloud Analytics Related Services
 - Wider Deep Learning Adoption

Analytics-Optimized Data

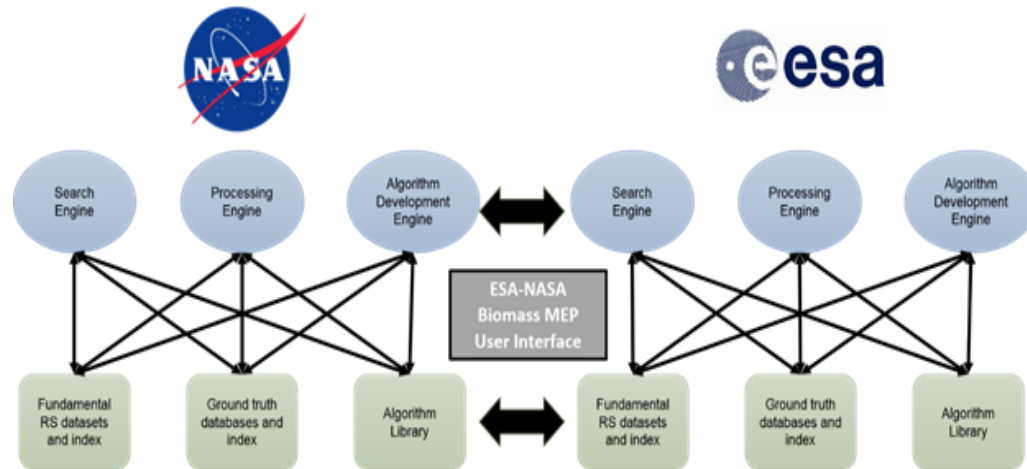
- Identified “Analytics Optimized Data Stores” (AODS) as a solution to Big Data Volume and Variety challenges
- AODS are data stored to minimize the need for data-wrangling for a large user community which enable fast access
- AODS are optimized, cost-effective storage structures enabling iterative queries relevant to users
- Building AODS to enable new analytic tools and services is viewed as a best path forward



Conceptual Architecture for Supporting Analytics



Multi-Mission Algorithm and Analysis Platform (MAAP)



Vision: The goal of the MAAP is to establish a collaboration framework between ESA and NASA to share data, science algorithms and compute resources in order to support scientific research conducted by NASA and ESA scientists.

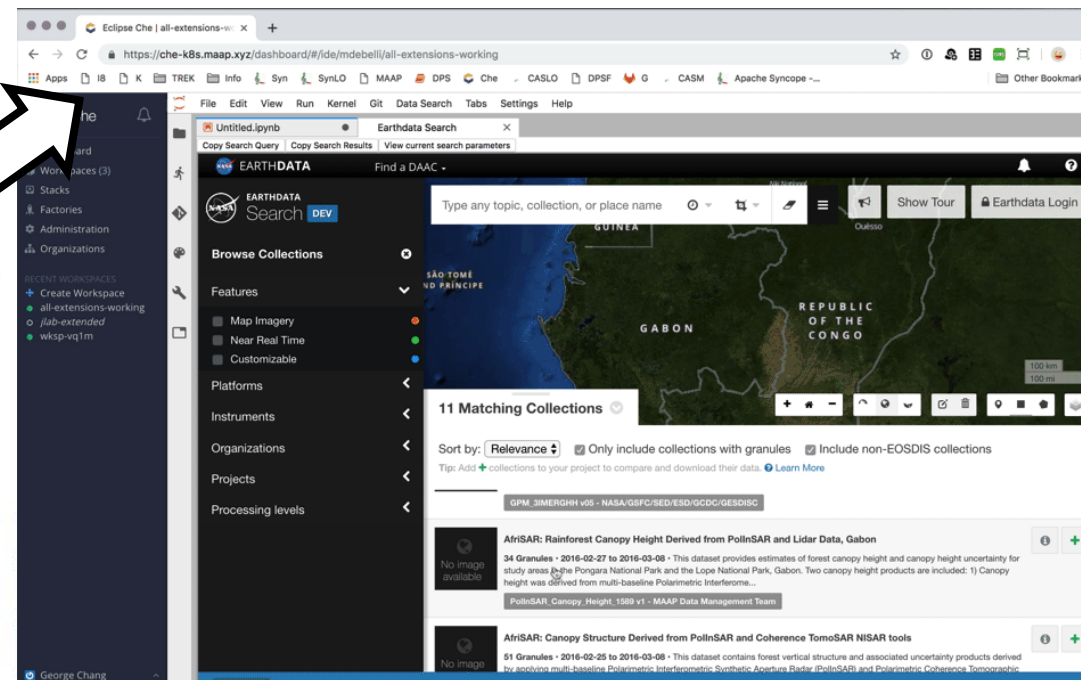
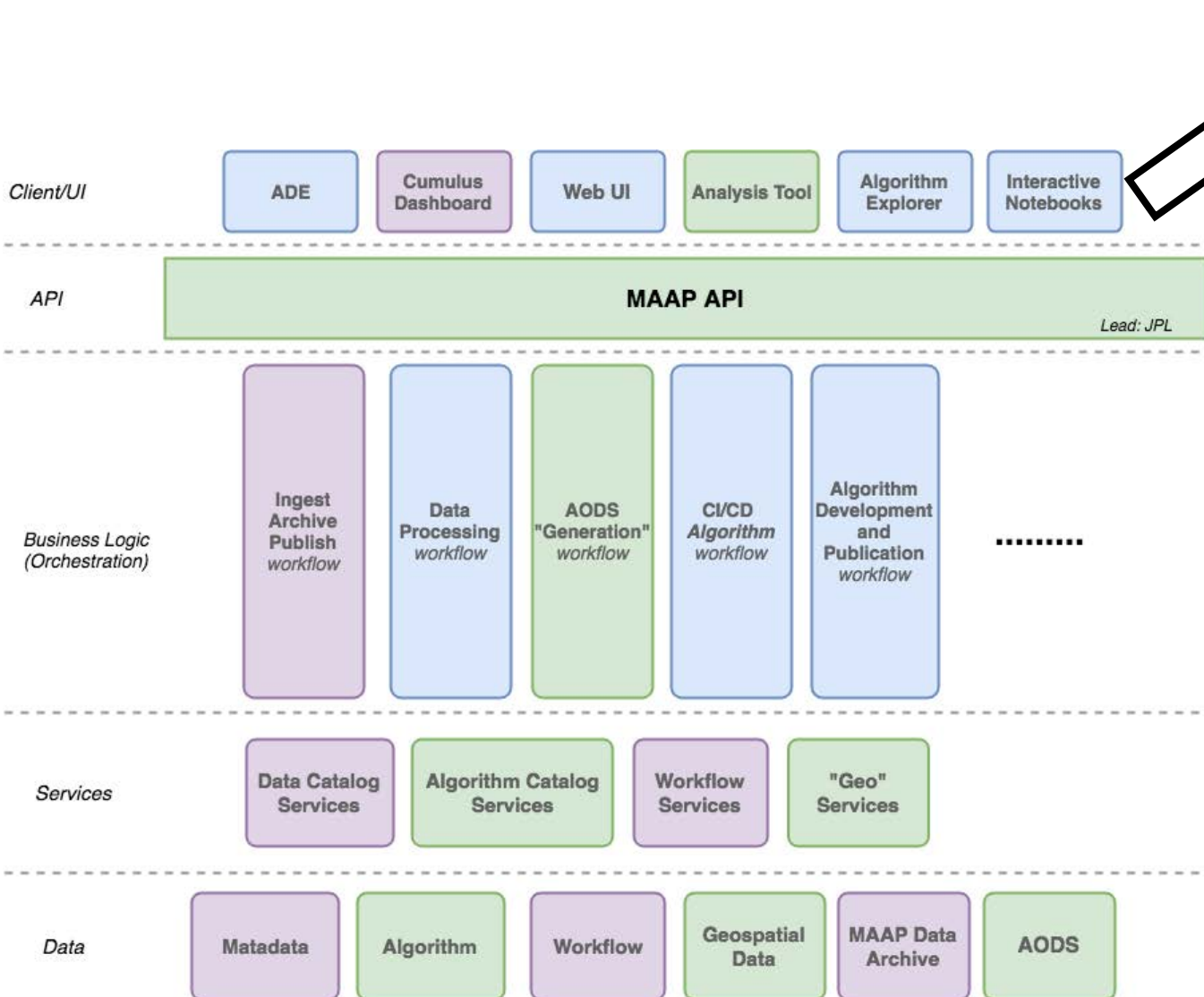
Outcomes

- Enable researchers to easily discover, **process**, visualize, and **analyze** large volumes of data from both ESA and NASA
- Provide tools and the supporting infrastructure needed to enable data comparison, analysis, evaluation, and generation

Deliverables

- Establish NASA MAAP data infrastructure including the use of CMR, MMT and Cumulus Workflow
- Ingest high priority datasets for pilot MAAP and curate metadata to high quality standards

MAAP Layered Architecture (Pilot)





Thank you!

Kevin Murphy ESDS Program PE, NASA HQ
Andy Mitchell ESDIS Project Manager, NASA GSFC
IMPACT Team, NASA MSFC