Supporting Open Science through Cloud-Based and Interoperable APIs for Data Access and Visualization: The Multi-Mission Algorithm and Analysis Platform (MAAP) Data System

Aimee Barciauskas (1), Kaylin Bugbee (2), George Chang (3), Laura Duncanson (4), Hook Hua (3), Aaron S Kaulfus (2), Alex Mandel (1)

(1) Development Seed, (2) NASA / University of Alabama in Huntsville, (3) Jet Propulsion Laboratory at California Institute of Technology, (4) University of Maryland

PRESENTED AT:

WHAT IS OPEN SCIENCE?

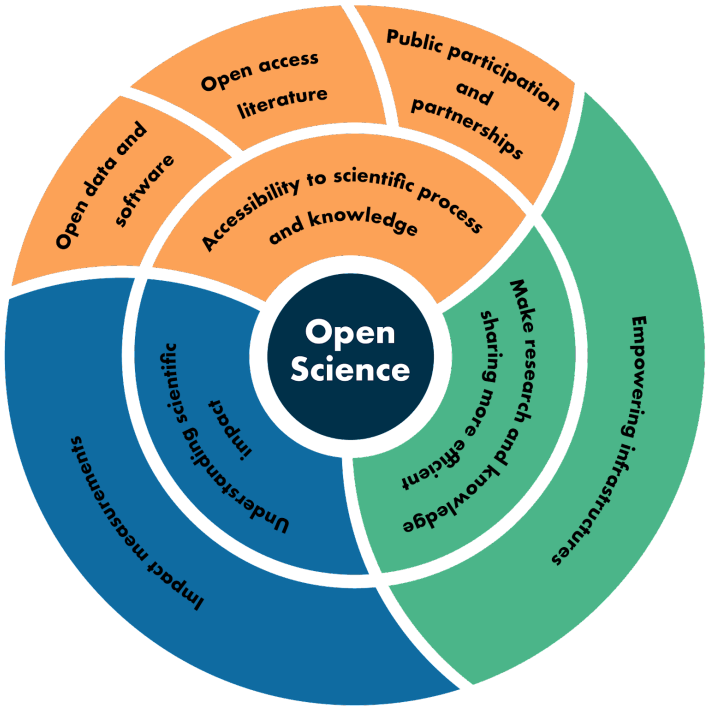


Image source: From Open Data to Open Science (2020), Rahul Ramachandran, Kaylin Bugbee, Kevin Murphy (https://doi.org/10.1029/2020EA001562)

Open science is **reproducible and transparent**, by anyone, anywhere.

Code and data are produced **collaboratively** and must not just be available but also **accessible** to anyone, through capacity building and web-based analysis environments.

Learn more about NASA's Transform to Open Science (TOPS) activities:

TOPS - highlights - public - latest - version
(https://docs.google.com/presentation/d/10zm_4dnHgLZE6d1HNjFJX89HfDDreVY970h8qVgeFFg/edit#slide=id.ge55f6c16de_0_98)

WHAT ARE KEY LESSONS FROM MAAP?

- **Reuse software and lessons from other systems.** Open science platforms are complex. They include systems for data storage to science storytelling and these components must form a cohesive whole to achieve the goals of open science.

- **Scientists' most critical needs are access to data and ability to share data they produce.** Focus on making it easy for scientists to access data and to share their data and results.

- **Make it easy to troubleshoot data processing.**

- **Make it easy for users to visualize and compare their products.**

- **When you can’t make it easy, write good documentation :).**

- **Focus on flexible development environments and data access methods.** Scientists quickly require different data and libraries than initially planned.

- **Capacity building and documentation are key.** It takes time and close collaboration to produce results. Development teams and scientists are learning how to build and use open science platforms.

- **Don’t move data around if you can help it.** Uniform and efficient access to high-value data is not a solved problem. Data storage is expensive and data has gravity. As datasets required by the biomass community are varied and stored across providers, data systems are evolving to provide uniform discovery and access across heterogeneous data localities and formats.

WHAT IS THE MAAP?

[VIDEO] https://res.cloudinary.com/amuze-interactive/video/upload/vc_auto/v1638749179/agu-fm2021/1E-D4-E2-0B-79-DA-30-85-AD-A7-57-4A-14-AC-DB-42/Video/icesat2_boreal_biomass_hyperwall_swsc9z.mp4

The MAAP is an open source science platform dedicated to the unique needs of sharing and processing data from relevant field, airborne and satellite missions.

MAAP Goal: The goal for the MAAP is to establish an Open Science collaboration framework between ESA and NASA to collaboratively share data, science algorithms and compute resources in order to foster and accelerate scientific research conducted by NASA and ESA scientists.

MAAP Objectives:

- An interoperable, set of shared resources integrated into a virtual collaborative platform in the cloud with special capabilities for seamless and efficient global Team Science activities.
- Full cloud-based sharing of appropriate large satellite and supporting data stores, as well as code and runtime images.
- Interoperability between cloud-based visualization, development and compute resources.
- Collaborative work environments with interoperability and full access to the above.
- Demonstration Use Case: Assessments of Global Above-Ground Biomass

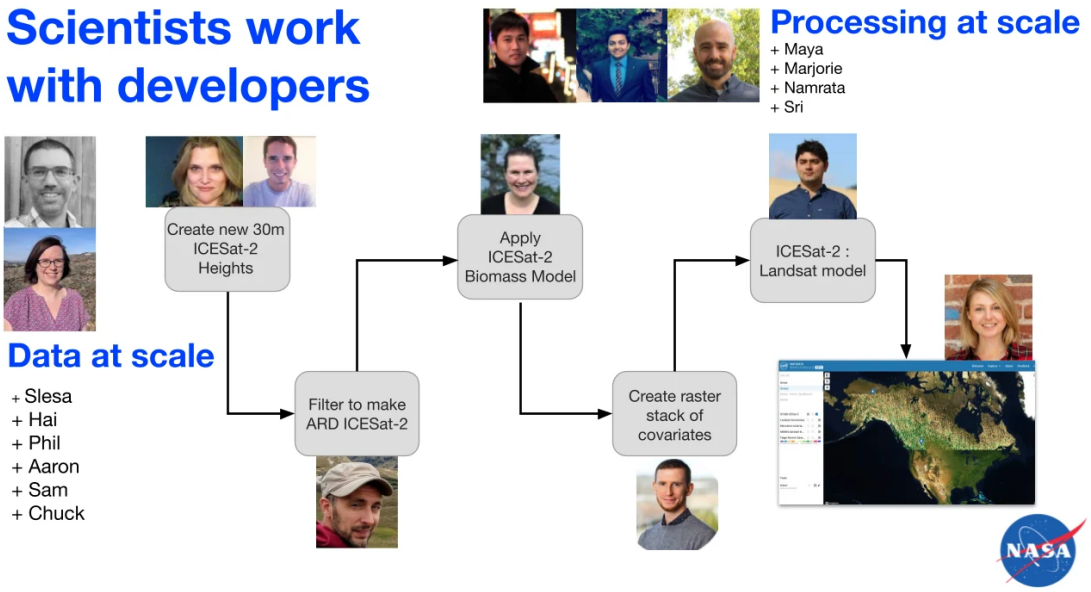
MAAP Activities:

- Weekly (Bi-weekly) ESA/NASA technical team meetings.
- Joint planning documents – e.g. Joint Schedule, Interoperability, Science Use Cases, High Level Requirements…
- Joint data, access and metadata standards.
- Deployment of full development, analysis and visualization environments
- Compatible and accessible cloud deployment of massive ESA/NASA LiDAR, SAR and other data sets
- Version 1 now completed and fully operational – public release was in October.
- Version 2 development and HECC integration now underway.

KEYS TO OPEN SCIENCE

Cross-functional Teams

The products developed on the MAAP are via collaboration of multiple development and science teams at ESA, NASA and the University of Maryland.



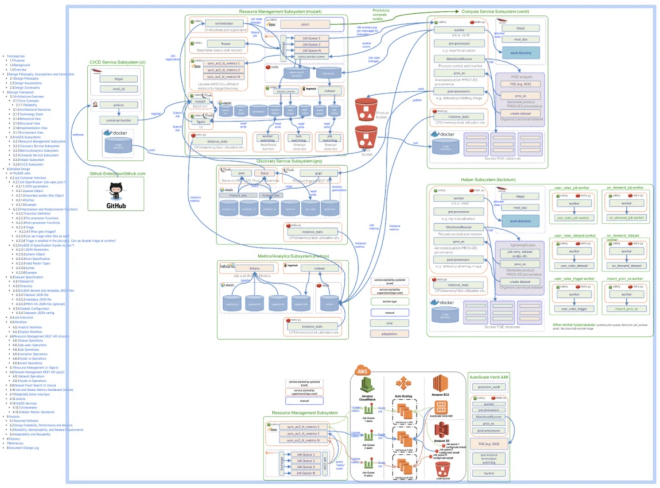
Access to Data

MAAP provides discovery of and access to high-value mission data via a distinct instance of NASA's Common Metadata Repository (CMR) and storage AWS S3.

Access to Compute (Scalable Data Procecssing)

The Hybrid Cloud Science Data System (HySDS) brings scalable compute by leveraging lower costs “spot fleets” while automatically handling the fault tolerances where compute are allowed to be taken away and/or fail while workloads can remain resilient to these faults.

HySDS-Core - Confluence (https://hysds-core.atlassian.net/wiki/spaces/HYS/overview)



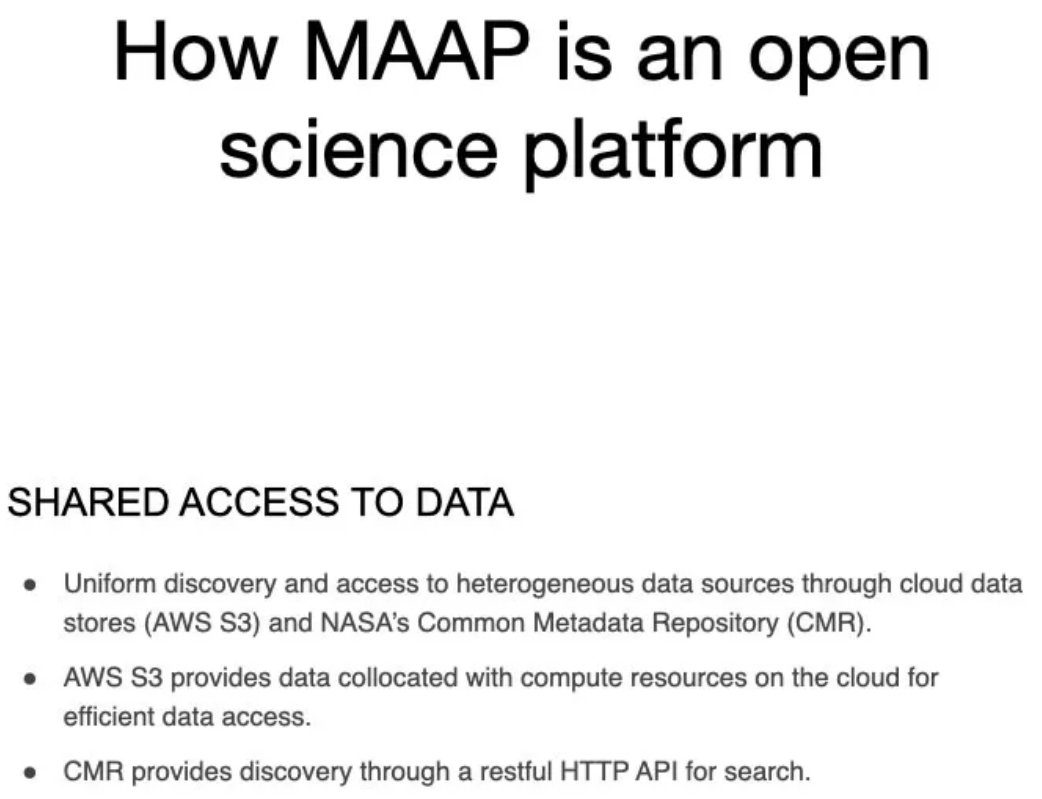
What HySDS looks like for developers

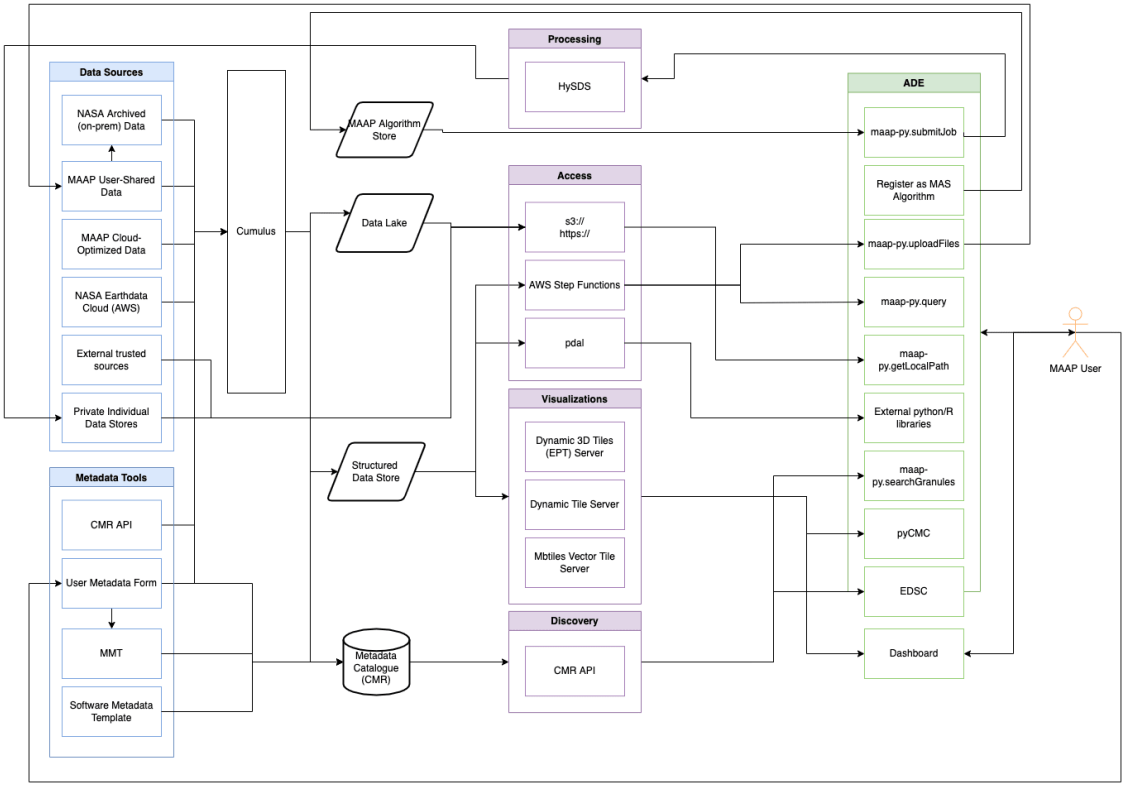
- fully automated and scalable system.
- write custom scripts or frameworks of any design or complexity.
- adaptable to a large range of projects or missions.

WATCH THE VIDEO

[VIDEO] https://res.cloudinary.com/amuze-interactive/video/upload/vc_auto/v1638751167/agu-fm2021/1E-D4-E2-0B-79-DA-30-85-AD-A7-57-4A-14-AC-DB-42/Video/MAAP_Final1_bvhieo.mp4

https://scimaap.net (https://scimaap.net)

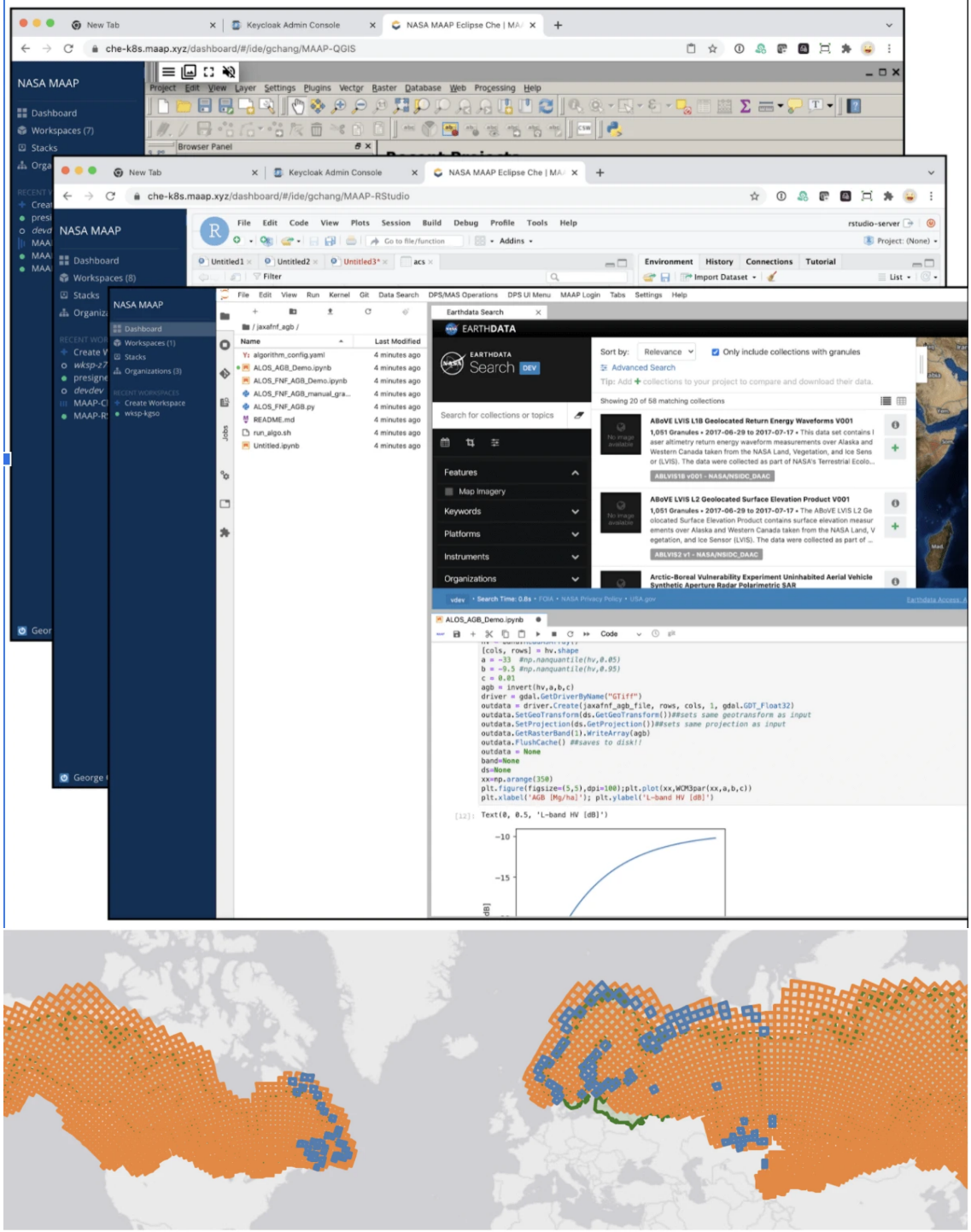




REPRODUCIBLE ANALYSIS

- Scalable and repeatable analysis through jupyter notebooks and a kubernetes cluster.
- Uses Eclipse Che, a cloud-based Kubernetes native IDE manager
  - Can run JupyterLab, RStudio, or other web-based IDEs for traditional programming languages
  - Customized workspaces
- MAAP built customized plugins and additions
  - Integrated Earthdata Search Client
  - Custom API to access MAAP functions
- Algorithm registration for large scale processing

Learn more about hysds, the framework behind MAAP's data processing system: HySDS-Core - Confluence.

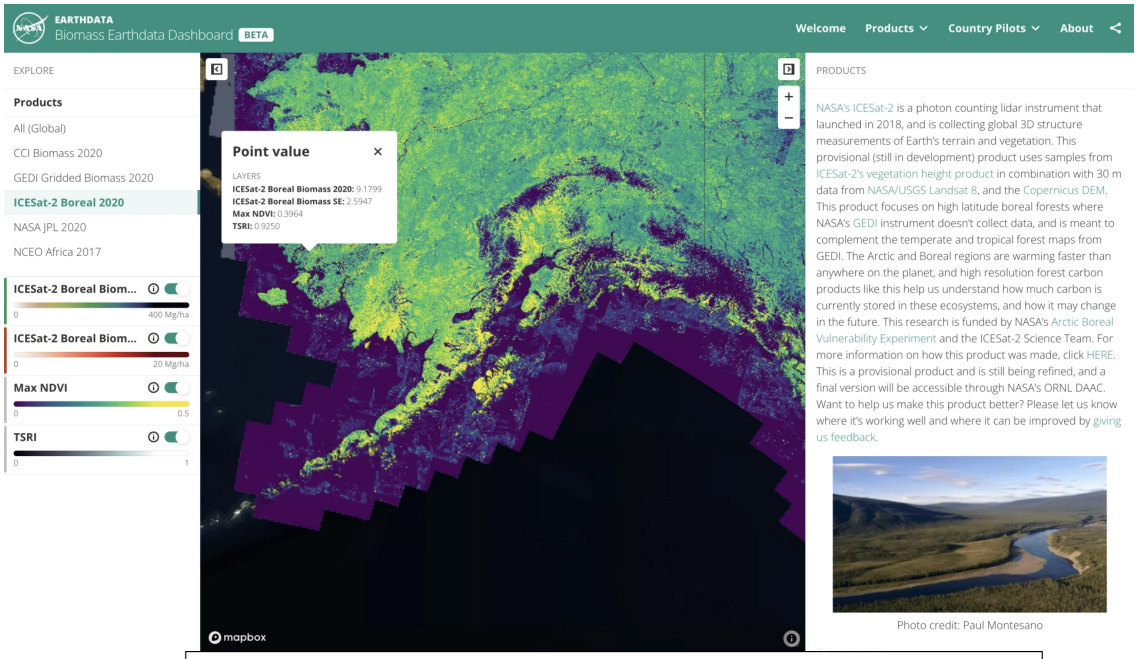


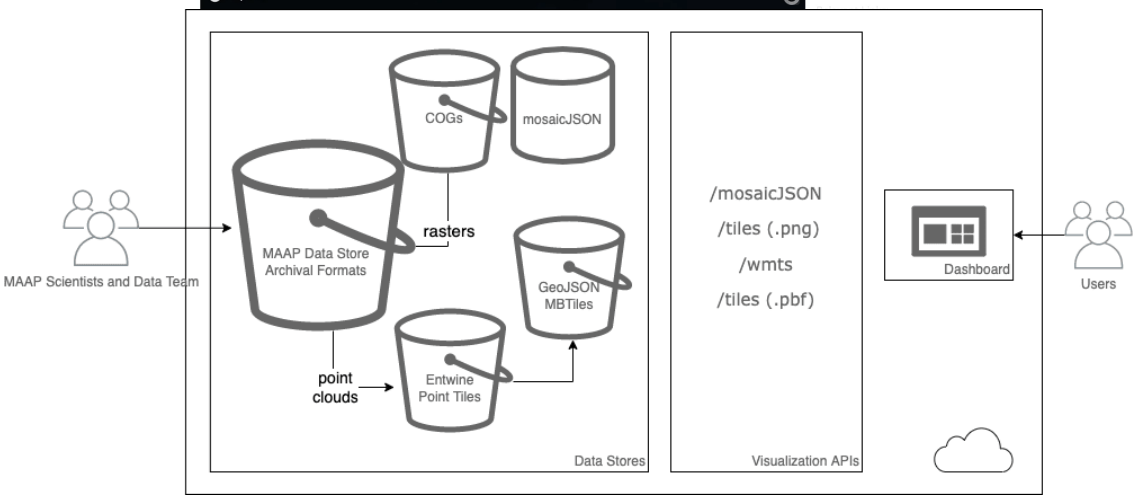
Screenshot: Calculating the Boreal Biomass Product. This is 4000 tiles which each take 4hrs on a single machine. We are running 1000 machines here in the cloud DPS.

How can new cloud-optimized data formats and services support open science?

Using just a single machine, no scientist can leverage the extent of massive and heterogeneous Earth data archives. MAAP and other open science platforms are adopting new data formats and access methods to provide subsetted access for real-time in-memory operations and cloud-based scalable compute environments.

MAAP has provided cloud-optimized access with formats such as Cloud-Optimized Geotiffs (COGs) and Entwine Point Tiles (EPT).





ABSTRACT

The aboveground biomass research community utilizes a diverse set of heterogeneous data for developing accurate estimation models. Data from field projects and airborne campaigns provide valuable in-situ data for calibration and validation while satellite remote sensing observations are required to create global biomass estimates. Furthermore, the computational requirement for global estimation requires high performance or scalable, dynamic infrastructures.

Innovative technological solutions lowering the barrier from data to scientific results are needed as both the volume of available data for refining biomass models and the scientific need for rapid processing increases.

The Joint ESA-NASA Multi-Mission Algorithm and Analysis Platform's (MAAP) is one such cloud-based solution that offers solutions to efficiently discovering, accessing, and processing at scale disparate data. This presentation will demonstrate the key components of the MAAP data system including the accessibility of data through interoperable APIs as well as subsetting and visualization of cloud-optimized data including lidar point clouds and raster datasets.

The data team strives to provide a high level of service for all datasets. Datasets must be processed and curated into these additional layers of service for subsetted data access and scalable visualization. The data team prioritizes this curation for datasets which have been identified as the highest value (such as the ICESat-2 ATL08 point cloud dataset and the SRTM Digital Elevation Model) and or the most challenging to use by the MAAP user working group.

This presentation will discuss the curation and management efforts required to support these services and user engagement efforts promoting open science to lowering the barrier to analysis.