BUILDING A DATA ECOSYSTEM: A NEW DATA STEWARDSHIP PARADIGM FOR THE MULTI-MISSION ALGORITHM AND ANALYSIS PLATFORM (MAAP)

Kaylin Bugbee[1], Manil Maskey[2], Aimee Barciauskas[3], Rahul Ramachandran[2], Aaron Kaulfus[1], Jeanné le Roux[1], Jeffrey Miller[1], Iksha Gurung[1], Amanda Whitehurst[4], Chris Lynnes[5]

[1]University of Alabama in Huntsville
[2]NASA Marshall Space Flight Center
[3]DevelopmentSeed
[4]ASRC Federal Technical Services
[5]NASA Goddard Space Flight Center

ABSTRACT

New adaptive approaches to Earth observation data stewardship need to be adopted in order to allow for higher data volumes, heterogeneous data and constantly evolving technologies. The data ecosystem approach to stewardship offers a viable solution to this need by placing an emphasis on the relationships between data, technologies and people. In this paper, we present the Joint ESA-NASA Multi-Mission Algorithm and Analysis Platform's (MAAP) creation of a data ecosystem to support global aboveground terrestrial carbon dynamics research. We present the components needed to support the MAAP data ecosystem along with two data stewardship workflows used in the MAAP and the development of extended metadata for MAAP.

Index Terms— Data stewardship, cloud technologies, metadata, data use, biomass

1. INTRODUCTION

Earth observation data volumes have grown exponentially as new platforms with more accurate sensors are launched each year [1]. Upcoming space borne missions, including ESA's BIOMASS mission and NASA/ISRO's NISAR mission, will offer unprecedented data about Earth but will also feature exponentially higher data volumes than any currently operating Earth observation mission. In addition, the heterogeneous nature of Earth observation data, which are collected from satellites, aircraft and ground stations at various resolutions, coverages and processing levels, makes analyzing these data a challenge for scientists.

In light of these ever growing data volumes, the diversity of Earth observation data and constantly evolving technologies, new approaches to data stewardship will need to be considered [2, 3]. While the traditional data stewardship approach of data publication is established and familiar, this approach may be too limiting to support these challenges. Instead, an approach is needed that views data stewardship as an evolving process that is flexible, collaborative and adaptive to support more effective data discovery and use [3]. The data ecosystem approach to stewardship addresses this adaptive need by emphasizing "the people and technologies collecting, handling, and using the data and the interactions between them" [3]. This data ecosystem approach is being leveraged by NASA and ESA, who are working together to develop the Multi-Mission Algorithm and Analysis Platform (MAAP). The MAAP will support the global aboveground terrestrial carbon dynamics research community by bringing together relevant data, algorithms and computing capabilities in a common environment.

In this paper, we describe NASA's creation of a data ecosystem for the MAAP. The MAAP's approach to data stewardship represents a new paradigm, where not only is the traditional data publication process replicated using new cloud-based technologies but data from disparate sources are also aggregated together to facilitate open data use. The MAAP data ecosystem is built on interactions between data providers, data curators, scientific subject matter experts who represent the target user community and technologists from both NASA and ESA. In the following sections, we describe the MAAP and the use cases leveraged to identify relevant data for the MAAP. Then we describe the MAAP data stewardship ecosystem, the MAAP data management workflows and the development of additional metadata information. Lastly, we conclude the paper with the goals and future direction of the MAAP.

2. MULTI-MISSION ALGORITHM AND ANALYSIS PLATFORM

NASA and ESA are collaborating on new approaches to lowering barriers to data use resulting from increased data volumes and to also reinforce open data policies. To meet this goal, NASA and ESA are jointly developing the MAAP to maximize the exploitation of Earth observation data from the BIOMASS, GEDI, and NISAR missions in order to improve the understanding of global aboveground terrestrial carbon dynamics. The MAAP will provide a platform where compute is collocated with data and where tools and algorithms are provided to support the biomass research community. The MAAP is leveraging cloud technologies to provide serverless computing, ease of use, the ability to scale up to extremely large datasets and the capability to collaborate across organizations. The MAAP will be developed in two phases: a pilot phase and a full phase. The pilot phase will demonstrate collaboration and basic capabilities and will focus on biomass relevant airborne and field campaign data while the full phase will focus on making data from the NISAR, GEDI and BIOMASS missions available in the MAAP.

In order to better understand the specific capabilities needed for the MAAP, sixteen broad example use cases were developed from discussions with biomass research scientists. A subset of these use cases were then identified in order to drive requirements and success criteria for the development of the pilot MAAP. Data needed to support the pilot MAAP were also identified by subject matter experts and prioritized with data needed to support the pilot use cases receiving the highest priority. Since the MAAP data repository provides access to data relevant to better understanding aboveground biomass research questions, the identification of data was not limited to data directly available from NASA or ESA. Instead, relevant open data from other Federal agencies and universities were also identified for inclusion in the pilot MAAP. The goal of this curated data collection is to make it easier for scientists to reuse data in research and algorithm development [4].

Over seventy relevant datasets were identified for the pilot phase. The curated datasets for pilot MAAP are heavily focused on the AfriSAR campaign, a field campaign that supported the upcoming BIOMASS, NISAR and GEDI missions and which collected "ground, airborne SAR and airborne Lidar data for the development and evaluation of forest structure and biomass retrieval algorithms" [5]. Supporting ancillary data, such as land cover products, SRTM DEMs and Landsat 7 data, were also identified for the pilot MAAP. Recognizing the limited development time of the pilot phase, data were prioritized to support the use cases with data essential for demonstrating the capabilities of the pilot MAAP receiving the highest priority.

3. MAAP DATA ECOSYSTEM

3.1. MAAP Data Ecosystem Components

The MAAP data team ensures the ongoing quality of the data and metadata provided in the pilot MAAP. The MAAP data team also supports the ingest and archive of data to the MAAP platform. In order to meet these goals, the MAAP data team deployed and is using a data stewardship system which reuses several open source software components developed by NASA's Earth Science Data and

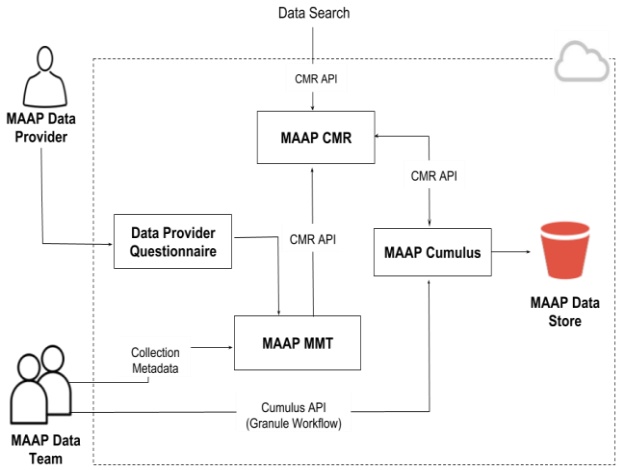

Fig. 1: The MAAP data ecosystem components and associated data management workflows.

Information System (ESDIS) Project. These open source components include (Fig. 1):

- The Common Metadata Repository (CMR): The CMR is an Earth science metadata repository for NASA's Earth Observing System Data and Information System (EOSDIS) system data.
- The Metadata Management Tool (MMT): The MMT is a web-based user interface that allows metadata authors to create and update CMR metadata records by using a data entry form based on metadata fields.
- Cumulus: Cumulus is a cloud-based framework for data ingest, archive, distribution, and management.
- CMR Application Programming Interface (API): The CMR API provides integration points for metadata ingest and search and is utilized by the MMT for collection metadata ingest.

Additional components, including a data provider questionnaire and a tool to crosswalk questionnaire information into the MMT, were developed by the MAAP data team to support data stewardship activities. These components are deployed in Amazon Web Services (AWS), the cloud provider for NASA MAAP. MAAP is leveraging native cloud services wherever possible to reduce the codebase of the platform.

The MAAP data stewardship system supports two different data management activities: traditional data

publication and data aggregation. These activities are described in more detail below.

3.2. MAAP Data Publication Workflow

Because some of the data identified for the pilot MAAP use cases are not currently available via a public archive, the MAAP data team needed to develop and implement a well-defined data publication process to ease the data sharing process for data providers and to ensure a high level of curation by the MAAP data team.

The MAAP data publication workflow replicates the traditional process of making "discrete, well-described data sets" [2] available in the pilot MAAP. This data publication process is similar to processes followed by NASA's Distributed Active Archive Centers (DAACs) and includes organizing and storing the data, developing collection and file level metadata for the data, and making the data discoverable to users in the pilot MAAP. A data provider questionnaire was developed by the MAAP data team to gather the essential information needed from the data provider in order to create collection level metadata, or information which describes the entire set of data files.

Once the provider has completed the data provider questionnaire and has uploaded the data to a MAAP AWS bucket, a converter tool crosswalks relevant information from the data provider questionnaire into the MMT. The MAAP data team then curates the dataset's collection level metadata using the data provider's information. If additional information is required in order to complete the metadata, the MAAP data team corresponds with the data provider until the metadata is as complete and accurate as possible. Once the data team is satisfied with the quality of the collection level metadata in draft form, the metadata record is published via the MMT to the MAAP CMR. While NASA's operational CMR was the logical choice for a metadata catalog for the MAAP, two requirements made deploying a MAAP instance of the CMR desirable. First, a shared metadata repository is a key component to interoperability between NASA's and ESA's data systems. A separate CMR was selected in order to facilitate ease of access to the CMR for ESA. Second, the pilot MAAP use cases identified a need for additional metadata curation beyond what was provided in existing metadata records. A separate deployment of the CMR enables changes to be made to metadata without impacting the original records.

Once the collection level metadata is published to the MAAP CMR, Cumulus transfers the data provider's dataset to the MAAP data store and the Cumulus workflow process publishes file level metadata to the MAAP CMR via the CMR API. As a part of the Cumulus workflow process, the MAAP data team identifies key file level metadata information that needs to be extracted from the files. A Cumulus workflow is then either created or reused that extracts and publishes the relevant information to the MAAP CMR. After the Cumulus process is complete, the MAAP data team checks both the collection and file level metadata for quality. Once these checks are complete, the metadata is available in the MAAP CMR and the data is made available via Amazon's S3 cloud service through more traditional data access methods such as https or sftp.

By using existing ESDIS software components, developing new components, and leveraging the commercial cloud, the MAAP data team successfully replicates the data publication process typically found at an on premise data center.

3.3. MAAP Data Aggregation Workflow

While data publication is an important activity for the MAAP data team, aggregating biomass relevant data and metadata into the MAAP is also an essential data management activity. Since the majority of the data identified by the use cases are data that are already publicly available at disparate archives, the pilot MAAP serves as a centralized and curated location for biomass relevant data. The pilot MAAP facilitates data discovery and use by aggregating the metadata into a single catalog and by aggregating the data into the MAAP data store.

For data that is already publicly available, the MAAP data team follows a workflow that is similar to the data publication workflow outlined above. Instead of asking a data provider to fill in the data provider questionnaire, the MAAP data team uses existing metadata to create the collection level metadata in the MMT and, when possible, to generate the file level metadata. To ensure metadata quality, the MAAP data team edits the provided metadata for any errors found and adds any additional metadata needed to support data discovery. Once these tasks are completed, the MAAP data team publishes the collection level record to the MAAP CMR. A Cumulus workflow ingests the data into the MAAP data store and publishes file level metadata. The MAAP data team validates all files were accurately ingested into the MAAP data store and also checks the collection and file level metadata for quality.

4. EXTENDED METADATA FOR UNIQUE SEARCH CRITERIA

As repository curators, the MAAP data team's goal is to ensure that "enough metadata is provided that others will be able to find and understand the data" [4]. Development of the use cases for the pilot MAAP have shown that the search needs of the biomass research community require more information than the existing metadata model provides. These needs included the ability to search for specific pieces of information that varied from platform and instrument type such as polarization, wavelength and heading for airborne SAR data, laser footprint diameter for lidar data and site

names for all field campaign data. In order to enable effective search and discovery for the biomass research community, the MAAP data team has developed over twenty additional metadata fields to support these unique search criteria. Examples of some of the information required to facilitate the requested searches are described in table 1.

The MAAP data team has not modified the existing set of fields provided by the CMR metadata model but has instead leveraged an existing field which allows for the description of unique characteristics of the data that extend beyond those defined in other metadata fields. This field is called 'Additional Attributes' and allows for a collection owner to include any number of additional metadata fields in both collection and file level metadata. The MAAP data team has written names, definitions and data types for each additional attribute needed and have added the fields to the relevant metadata. While some of the identified additional metadata may be described in other parts of the metadata model, the MAAP data team opted to leverage additional attributes for the pilot phase. This choice was made in order to facilitate ease of collaboration and metadata interoperability between NASA and ESA. In addition, some attributes in the metadata model cannot be searched via the CMR API. Therefore, leveraging additional attributes makes searching via the CMR API possible.

Platform/Instrument Type	**Search Request Criteria**	**Example Values**
Airborne & Satellite SAR	Polarization	HH, HV, VH, VV
Airborne SAR and Lidar	Flight Number	16008
Airborne SAR & Lidar, Terrestrial Lidar, In situ Measurements	Research site name	Lope National Park

Table 1: Examples of information needed to search for biomass data in the MAAP CMR.

5. CONCLUSIONS AND FUTURE WORK

The NASA MAAP data team has created a data ecosystem that establishes a new paradigm for data stewardship including the replication of the traditional data publication process in the cloud and the aggregation of relevant data into a centralized, cloud-based location. In addition, the MAAP data team has explored new ways for curating metadata to meet the unique search needs of a specific Earth observation research community. As work continues on the pilot MAAP, the NASA MAAP data team will continue to be pathfinders for novel solutions to enhancing the MAAP data ecosystem. Future work includes the creation of a prototype metadata model to describe software and algorithms, the development of Analytics Optimized Data Store (AODS) to support data expeditions in the cloud, and the implementation of data sharing by users in the pilot MAAP. This future work will not only support the development goals of the MAAP but will also support the long term goal of establishing the MAAP as a well curated repository that is recognized by the community as a trustworthy location for uploading data and conducting research [4]. Additionally, the development of operations concepts is planned for the new data stewardship paradigm.

Lastly, in order to make biomass relevant data more discoverable and usable in the pilot MAAP, NASA and ESA are collaborating together to make metadata and data more interoperable across organizations. To support this interoperability for the pilot phase, ESA plans to use the MAAP CMR to ensure that all relevant metadata are provided in a centralized location. Additionally, ESA plans to use the MMT to curate collection level metadata to support metadata interoperability. Since additional metadata is key to search and discovery, NASA and ESA are working together to identify and refine the additional metadata required to support the biomass research community needs and to implement the additional metadata in the MAAP CMR. This focus on collaboration and interoperability will ensure that the MAAP lowers barriers to both the discovery and use of key data needed to increase scientific discoveries in the aboveground terrestrial carbon dynamics research community.

6. REFERENCES

[1] S. Nativi, P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia and O. Ochiai, "Big Data challenges in building the Global Earth Observation System of Systems," *Environmental Modelling & Software*, vol. 68, pp. 1 - 26, Jun. 2015.

[2] M. Parsons and P. Fox, "Is Data Publication the Right Metaphor?" *Data Science Journal,* vol. 12, pp 32 – 46, Feb. 2013.

[3] M. Parsons, ∅ Godoy, E. LeDrew, T. de Bruin, B. Danis, S. Tomlinson and D. Carlson, "A conceptual framework for managing very diverse data for complex, interdisciplinary science," *Journal of Information Science,* vol. 37, no. 6, pp. 555 – 569, Oct. 2011.

[4] Y. Gil, C. David, I. Demir, B. Essawy, R. Fulweiler, J. Goodall, L. Karlstrom, H. Lee, H. Millis, J. Oh, S. Pierce, A. Pope, M. Tzeng, S. Villamizar and X. Yu, "Towards the Geoscience Paper of the Future: Best Practices for documenting and sharing research from data to software to provenance," *Earth and Space Science,* vol. 3, no. 10, pp. 388 – 415, Oct. 2016.

[5] L. Fatoyinbo, N. Pinto, M. Hofton, M. Simard, B. Blair, S. Saatchi, Y. Lou, R. Dubayah, S. Hensley, J. Armston, L. Duncanson and M. Lavalle, "The 2016 NASA AfriSAR campaign: Airborne SAR and Lidar measurements of tropical forest structure and biomass in support of future satellite missions," in *Proc. International Geoscience and Remote Sensing Symposium (IGARSS),* 2017.