Data Stewardship Practices for Earth Observation Transient and Optimized Analysis Platforms

Kaylin Bugbee[1], Rahul Ramachandran[1], Ge Peng[2] and Aaron Kaulfus[1]

[1]NASA Marshall Space Flight Center, [2]University of Alabama in Huntsville

IMPACT
Interagency Implementation and Advanced Concepts Team

Introduction

Science is evolving and changing, with several factors driving the change [1]. First, technological advances have enabled new scientific workflows, easier sharing of data and code among scientists, new sensors for collecting observations and new techniques for analyzing data. These technological advances have, in turn, increased the volume and heterogeneity of data available to scientists [2]. These data may originate from new sensors, new modeling and analysis techniques or from new sources like citizen science activities. Lastly, as more scientists and organizations embrace the idea of open science, increasing amounts of data, software and documentation will be available for analysis [1].

These advances in science are driving the development of specialized, transient analysis platforms in the cloud or using high-performance computing (HPC) solutions. These new platforms require a different approach to the data stewardship lifecycle - data and information needs to be managed in a standardized and uniform manner throughout the platform but are not subject to some of the traditional curation approaches. In this poster, we describe what a transient optimized platforms, the need for evolving curation approaches in these platforms and our experiences managing data in a transient optimized platform.

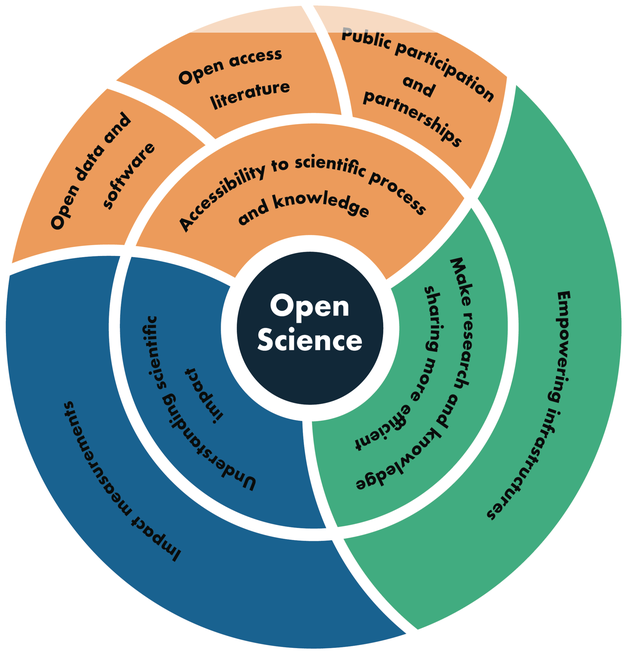

Figure 1: The open science concept includes three broad focus areas: (1) Accessibility to the scientific process and knowledge (2) Making research and knowledge sharing more efficient (3) Understanding scientific impact. [1]

What are transient and optimized analysis platforms?

- Provides scalable, flexible storage and compute resources for large-scale, efficient scientific analyses [3]
- Instantiated to address different scientific needs
- Some are general purpose platforms while others are specialized platforms that focus on supporting scientists conducting research on specific domain or problem areas with heterogeneous data types and customizations of tools and data
- Not meant to be a permanent archive of record and are not subject to the commitment of a long-term archive. Have a pre-defined, or transient, life span.
- Focused on providing improved analysis capabilities and developing a flexible environment for exploring the effects of cloud computing and scientific analysis at scale on open data access and cost
- Science is still being conducted in these platforms so there is a need to support reproducibility, data quality information and other key open science aspects

Data Curation Approaches

The goal of data curation and the data curation lifecycle is to ensure the continuity of digital objects [4] such as data, software, code and documentation. The meaning of 'continuity' varies from organization to organization, resulting in a spectrum of data curation methodologies addressing a variety of needs. Many traditional archives simply focus on preserving, maintaining and distributing data while some operational, domain specific archives have recognized the need to move beyond the basic data curation services. These operational archives focus on providing enhanced services for select data, guided by a level of service framework, to improve data access and use. Transient and optimized analysis platforms, on the other hand, consider the entire ecosystem of data storage, pipelines, tools and services within a platform. The platform is considered holistically instead of siloed, making a cohesive governance plan possible. This approach may more closely resemble information governance plans often adopted for enterprise data platforms as seen in industry. In the end, transient, optimized analysis platforms approach data curation as living and evolving activity like maintaining a vegetable garden as opposed to preserving data like 'a time capsule or Egyptian tomb' [5].

Multi-Mission Algorithm and Analysis Platform (MAAP)

The MAAP is an open science, collaborative platform dedicated to the unique needs of sharing and processing data from relevant field, airborne and satellite measurements related to ESA and NASA missions. The MAAP is jointly managed by ESA and NASA. The science focus of the MAAP is to improve the understanding of global terrestrial carbon dynamics and to support algorithm development within that research community.

The MAAP has several unique data curation needs. These needs include:

- Curating domain-specific metadata to enable specialized search. This specialized curation has focused on providing metadata unique to sensing techniques, such as SAR and lidar, and metadata for specific field campaigns. Providing this metadata required an expansion of the existing metadata model being used.
- Providing cloud-optimized datasets to enable large-scale and efficient analyses. This includes generating the datasets from the archival copy, documenting the generation process to address data quality and provenance and making these datasets discoverable by distributing them within the platform as unique collections.

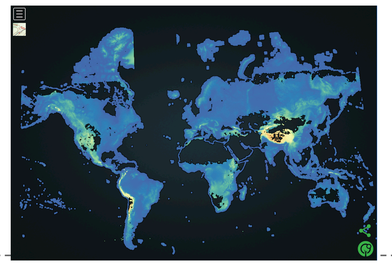

Figure 2: Sample visualization of the cloud-optimized global ATL08 digital elevation model entwine point tile store. The point tile store was generated on the MAAP and is available to users for analysis.

References

1. R. Ramachandran, K. Bugbee, and K. Murphy. "From Open Data to Open Science," Earth and Space Science Open Archive, 2020. https://doi.org/10.1002/essoar.10505011.1.
2. Yao X, Li G, Xia J, Ben J, Cao Q, Zhao L, Ma Y, Zhang L, Zhu D. Enabling the Big Earth Observation Data via Cloud Computing and DGGS: Opportunities and Challenges. Remote Sensing. 2020; 12(1):62. https://doi.org/10.3390/rs12010062
3. Gomes VCF, Queiroz GR, Ferreira KR. An Overview of Platforms for Big Earth Observation Data Management and Analysis. Remote Sensing. 2020; 12(8):1253. https://doi.org/10.3390/rs12081253
4. S. Higgins. "The DCC Curation Lifecycle Model," The International Journal of Digital Curation. 2008; 1(3).
5. B. Heidron. "The Emerging Role of Libraries in Data Curation and E-science," The Journal of Library Administration. 2011. DOI: 10.1080/01930826.2011.601269

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

NASA