NASA's Earth Observing Data and Information System – Near-Term Challenges

Jeanne Behnke[1], Andrew Mitchell[1] and Hampapuram Ramapriyan[1, 2]
[1]NASA Goddard Space Flight Center
[2]Science Systems and Applications, Inc.
Corresponding author's e-mail address: Jeanne.Behnke@nasa.gov

Abstract

NASA's Earth Observing System Data and Information System (EOSDIS) has been a central component of the NASA Earth observation program since the 1990's. EOSDIS manages data covering a wide range of Earth science disciplines including cryosphere, land cover change, polar processes, field campaigns, ocean surface, digital elevation, atmosphere dynamics and composition, and inter-disciplinary research, and many others. One of the key components of EOSDIS is a set of twelve discipline-based Distributed Active Archive Centers (DAACs) distributed across the United States. Managed by NASA's Earth Science Data and Information System (ESDIS) Project at Goddard Space Flight Center, these DAACs serve over 3 million users globally. The ESDIS Project provides the infrastructure support for EOSDIS, which includes other components such as the Science Investigator-led Processing systems (SIPS), common metadata and metrics management systems, specialized network systems, standards management, and centralized support for use of commercial cloud capabilities. Given the long-term requirements, and the rapid pace of information technology and changing expectations of the user community, EOSDIS has evolved continually over the past three decades. However, many challenges remain. Challenges addressed in this paper include: growing volume and variety, achieving consistency across a diverse set of data producers, managing information about a large number of datasets, migration to a cloud computing environment, optimizing data discovery and access, incorporating user feedback from a diverse community, keeping metadata updated as data collections grow and age, and ensuring that all the content needed for understanding datasets by future users is identified and preserved.

Keywords: Data systems, Earth science, Remote sensing, Big data, Data discovery, Data access, Data preservation, Metadata

Introduction

NASA's Earth Observing System (EOS) Data and Information System (EOSDIS) has been a central component of the NASA Earth observation program since the 1990's. The data collected by NASA represent a significant public investment in research. Consequently, NASA developed a free, open and non-discriminatory policy consistent with existing international policies to maximize access to data. EOSDIS manages data covering a wide range of Earth science disciplines. The data managed by EOSDIS include observations from instruments on board satellites and aircraft, in situ observations from field campaigns, and digital products derived from such observations. The EOSDIS is comprised of partnerships among NASA Centers, other US agencies and academia that process and disseminate remote sensing and in situ Earth science data. One of the key components of EOSDIS is a set of twelve discipline-based Distributed Active Archive Centers (DAACs). Because of their active role in NASA mission science and with the science community, they perform many tasks beyond basic data stewardship, representing a distinct departure from typical data archives. They are collocated with scientific expertise in their respective Earth science disciplines. Managed by NASA's Earth Science Data and Information System (ESDIS) Project at Goddard Space Flight Center and geographically distributed across the United States, these DAACs serve over 3 million users globally. The ESDIS Project provides the infrastructure for EOSDIS including other components such as the Science Investigator-led Processing systems (SIPS), common metadata and metrics management systems, specialized network and security systems, standards management, and centralized support for use of commercial cloud capabilities. Given the long-term requirements, and the rapid pace of information technology and changing expectations of the user community, the ESDIS Project has had to evolve EOSDIS continually over the past three decades to address many challenges. The purpose of this paper is to describe some of the key challenges and the approaches being taken to address them. In the era of big data, it is important to consider emergent issues in light of the ongoing challenges that science archives like EOSDIS have addressed and are continuing to address.

Challenges

As a long-lived system that manages data from many diverse sources and serves a multi-disciplinary user community, EOSDIS faces many challenges. These challenges can be grouped into three main categories: 1. Managing volume and variety; 2. Enabling data discovery and access; and 3. Incorporating user feedback and concerns. These challenges are discussed in the three subsections below.

Managing volume and variety

Back in the 1990s, when EOSDIS was conceived it was understood that it would always be a growing collection of Earth science datasets. It started with very small collections that NASA had funded at various locations, which became the DAACs of the EOSDIS. The NASA EOS program was planned to consist of several multi-instrument platforms that would collect data continuously. From the management and funding perspective, it makes sense to have a single system that manages multi-mission operations, as opposed to the old model of creating a new processing/archiving system with each mission. Since its inception, EOSDIS has added new missions to the Earth science collection expanding the variety and volume every year. With each orbit, instruments continue to acquire data adding to the collection. However, the data also grows as scientists improve the measurements from the instruments deriving new parameters and products. Data formats that were chosen at launch must adopt to meet new standards and feed new software applications reliant on improved metadata. In the 1990s, staff at the ESDIS Project had a difficult time convincing scientist data providers of the value of metadata – assuring them that the most metadata they would have to insert into the complex Hierarchical Data Format – HDF format would be no more than 18 individual fields. Today, metadata is a ubiquitous word – everyone understands the value of it and the EOSDIS metadata model has grown to cover not only data, but services, identifiers, humanizers and so on.

At this time, EOSDIS has over 400 million granules identified in its repository from over 7,000 data collections. The number of providers who are allowed to load data into and delete data from the repository is controlled. The EOSDIS Common Metadata Repository software to manage this is carefully maintained, but open source versions of the software are available along with programming interfaces that allow anyone to access the repository. The ESDIS Project provides an infrastructural software system, Earthdata Search, as a user interface to the repository. Because of the diversification by discipline, the workload in maintaining the EOSDIS collection is shared by the DAACs. Vigilance is still needed as inconsistencies across the collected information become more readily apparent in Earthdata Search and other user interface applications. One way ESDIS manages inconsistencies is to establish an independent review committee composed of metadata professionals. These professionals have scrubbed through the DAAC collections, focusing on metadata, to create targeted reports of errors, misspellings, inconsistencies, etc. This two-year task is expected to improve the user experience in searching through the EOSDIS data collection.

As has been the case over the history of EOSDIS, the diversity of science data producers contributing to the variety and volume of the collection continues to present challenges. Physical storage and hardware challenges are always expected as the collection grows. However, the challenge presented by the diversity of data producers is inescapable. Although we have required standards for data and metadata, like the HDF and ISO19115 (ESDIS 2017) , we do not precisely control the way the standard is implemented by a particular science instrument team or SIPS. This challenge means that with each organization-wide system change, whether for new versions or transferring to new technology, each dataset must be handled individually. An additional challenge now is that several of the original EOS instruments are at end of their life. In the old days, data would be written to tape and racked and users could request the data but would have to figure out how to read the tapes. Since all data are now online and available, the data are easier to maintain and re-version, but we may no longer have final versions of data. These heritage datasets will always need to be maintained at the DAACs and the ESDIS Project plans ahead to keep them updated to the latest data and metadata standards. In the case of ICESat (2003-2010) data, the DAAC archived the final version of the data several years ago but continues to update the metadata to make it discoverable by users. However, researchers who have better calibration data have reprocessed new versions of the entire dataset and it is difficult for the archive staff to know what to do about these newer datasets. Procedures need to be established for deciding whether they should replace the older versions and distributed by the archive to users.

EOSDIS, like many other space data archives, is looking at the use of the commercial cloud as the next avenue for data storage and services. Instead of managing in-house hardware systems at the DAACs, the use of cloud systems is very appealing. Several prototyping tasks have been undertaken to gauge the effort of managing data in commercial cloud structures (McInerney 2017). The chief effort is to build an infrastructure that allows all of the EOSDIS components to work in a controlled fashion on the various cloud platforms. One challenge is the effort to

make certain that we understand the security aspects associated with the use of a commercial system. We are working on a specific security plan that would identify all risks and contingencies associated with the use of a commercial entity. Another issue is the use of various networks to access the cloud systems. Improper use of the networks could increase the cost of cloud use significantly. Several test efforts are ongoing to document various traffic patterns and usage. In addition, we have prototyped the development of a system, based on existing processes, for ingest and archive of data. This system is undergoing functional and performance tests to work out the many issues that have been encountered. However, one of the greatest challenges will be managing the overall cost of using the cloud by the various components of EOSDIS. Developing the processes for such management are ongoing and proving to be problematic but not insurmountable. We expect that by using the commercial cloud as a platform, the advantages to the user community will be myriad. Researchers will be able to gain new insights into the data and users will enable new applications, which ultimately is the end goal for the big data era.

Enabling Data Discovery and Access

A continuing challenge is to provide users with just the data they need. Typically users search for data using keywords as well as spatial and temporal constraints. In EOSDIS, with thousands of datasets, typical queries from users may result in hundreds of hits meeting their criteria. Ensuring that the most relevant of the datasets appear first in the results list is crucial to users. The obvious steps one can take towards increasing search relevance are ranking the datasets based on spatial and temporal relevance. Also, ranking newer versions higher, and applying information about community usage of datasets (e.g., through automated analyses of scientific literature) for ranking are useful steps. Observation of the real usage of the Earthdata search capability in EOSDIS and characterizing the search and access will also help in continuous improvements to data discovery.

In the case of data that can be represented as images, it is beneficial for users to be able to visualize them and select the data that they want to download and analyze. Enabling this for large volume datasets is a challenge that we have successfully addressed through its Global Imagery Browse System (GIBS) and the WorldView client software (Murphy et al. 2015 and David et al. 2015). The GIBS consists of a database of images stored in a hierarchical manner to enable rapid access to data at multiple resolutions. The WorldView client takes advantage of this data structure and enables users converge within a few seconds on their area of interest at the highest resolution offered by the dataset.

The access to data by users has changed significantly over the last two decades. In the 2000's, EOSDIS data were stored in robotic tape silos. Users would discover what they needed and place on-line orders for data from the respective DAACs. The DAACs would copy data to media and mail them to the users or stage the data on disk and email users so that they could download the data. With the move starting in 2006 to online storage, users now select data granules (files) that meet their search criteria and are provided with the URLs, which they can use to download the granules. The on-line storage also has enabled the users to request services such as subsetting, reformatting and reprojection conveniently prior to downloading the data. However, as the volume of data is expected to increase significantly in the near future, new challenges arise.

Providing to users data that are ready for ingest into algorithms and for analysis saves them considerable amount of traditional preparatory work, such as downloading large amounts of data, subsetting, reprojection, mosaicking, etc. This idea of "analysis-ready data" is becoming more popular recently, especially with respect to Landsat data (USGS 2018). Of course, to extend this idea to all the Earth science disciplines is a challenge due to the differences in the way different science disciplines deal with data. Defining analysis-ready data for different disciplines and preparing the data to meet their diverse needs would take significant effort, especially in a well-established system such as EOSDIS with hundreds of millions of data files requiring reorganization. The next step is to carefully evaluate typical use cases in different disciplines and prioritize implementation efforts. Also, the large and increasing volumes of data make it impractical for users to download them into their own systems for analysis. Near-archive analysis capabilities, as in the case of archiving data in a commercial cloud environment, will alleviate this problem significantly. The challenges of security and managing costs in the cloud environment are real, and are being addressed as described earlier.

Another challenge in this area is ensuring access to data decades into the future. The data and derived products from NASA's missions are a valuable asset resulting in many important scientific discoveries and influential findings. Therefore, they need to be preserved so that future users are able to discover, access, read, understand and reuse them. Future users should be able to verify, reproduce or question the science as necessary without having access to

the science teams that produced the products. The contents needed to be preserved with the data can be referred to as associated knowledge. ESDIS developed a "Preservation Content Specification" that identifies the classes of content that need to be preserved (NASA, 2011). Similar efforts have been documented by the European Space Agency and the Committee on Earth Observing Satellites (CEOS) Working Group on Information systems and Services (WGISS) (CEOS, 2015) Since then having educated our components, DAACs and SIPS, on preservation of content, especially with regard to instruments at end-of-life. It is important that the broader community also consider the serious issues of long-term data archive and accessibility, so we have worked to encourage the universal comment on ways to preserve data along with adoption of these practices.

Incorporating User Feedback and Concerns

As a system that serves a diverse global community of over 3 million users, EOSDIS receives feedback from them in several different ways. Responding to the diversity of the feedback is a challenge. Users can provide direct feedback including suggestions, problem reports or questions on the webpage http://earthdata.nasa.gov. Each of the DAACs has a user services team responsible for analyzing applicable user feedback and responding to requests for help. Also, each of the DAACs has a user working group (UWG) consisting of science and applications users representing the DAAC's specific discipline(s). The UWGs meet periodically with the ESDIS and DAAC staff members to review the data holdings, tools and services offered by the DAACs and provide advice on priorities and future plans. The ESDIS Project employs an independent organization to conduct an annual survey of users to derive the "American Customer Satisfaction Index (ACSI)". While the ACSI is a number indicating how satisfied the users are, the survey also includes several questions for which users provide free-form answers. The ESDIS Project and DAACs analyze these answers for suggestions for system improvement. In addition, focused efforts have been made within NASA's Earth Science Data System Working Groups (ESDSWG) for user needs assessment.

A related challenge is a concern by users regarding privacy while the system requires them to be registered in order to obtain most of the data and services. As a system that manages data from NASA as well as other non-U.S. partners, EOSDIS must comply with different rules regarding access restrictions and privacy policies regarding collection of information about data users. NASA has had a free and open data policy for Earth science data since the beginning of the EOS Program in 1990. However, working under agreements for archiving and distributing data from international partners, NASA complied with more restrictive policies regarding charging for data and requiring users to be registered and authorized to obtain the partners' data. Until 2012 NASA did not have a registration system for users to access the data from NASA missions. NASA's "earthdata login" is now used for registering users, but with minimal information needed for registration so that more accurate metrics are collected about numbers and organizations of users, and users can be contacted about new datasets and features offered by EOSDIS. The need for better metrics and services to users is balanced relative to privacy rules.

Conclusions

As a long-lived data system, EOSDIS has faced a number of technological, organizational sociological challenges over the past two decades. It continues to evolve in response to such challenges, but the challenges are not unique to EOSDIS. By sharing our issues and solutions, we look forward to discussions of state-of-the-art solutions and novel data services used in other scientific data archives. It is clear that we are all stepping into the big data era.

Acknowledgements

Jeanne Behnke and Andrew Mitchell contributed to this paper as a part of their duties as employees of NASA. Hampapuram Ramapriyan was supported by NASA contract NNG15HQ01C with Science Systems and Applications, Inc.

References

CEOS 2015 Earth Observation Preserved Data Set Content (PDSC), http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf (last accessed April 26, 2018)

Davies, D. et al. (2015) "The use of NASA LANCE imagery and data for near real-time applications", Time-sensitive remote sensing, Springer, New York, NY, 2015. DOI: https://doi.org/10.1007/978-1-4939-2602-2_11

ESDIS 2017 Standards, Requirements and References, https://earthdata.nasa.gov/about/system-performance, (last accessed April 26, 2018)

McInerney, M 2017 EOSDIS Cloud Evolution, https://earthdata.nasa.gov/about/eosdis-cloud-evolution, (last accessed April 26, 2018)

Murphy, K. J. et al. 2015 LANCE, NASA's Land, Atmosphere Near Real-Time Capability for EOS, *Time-Sensitive Remote Sensing*, Springer, New York, NY, 2015. DOI: https://doi.org/10.1007/978-1-4939-2602-2_8

NASA 2011 NASA ES Data Preservation Content Spec (423-SPEC-001, Nov 2011), http://earthdata.nasa.gov/sites/default/files/field/document/NASA_ESD_Preservation_Spec.pdf (last accessed April 26, 2018)

USGS 2018 U.S. Landsat Analysis Ready Data, https://landsat.usgs.gov/ard (last accessed April 26, 2018)