Are SMD's Science Data Archives Ready to Meet Future Challenges?

Background

The NASA Science Mission Directorate (SMD) supports a wide variety of facilities and operations dedicated to producing and serving mission data to the research community. As part of its review of NASA big data activities, the BDTF reviewed the status of all the major data center activities. These included ESDIS, HEASARC, MAST, PDS, SDAC, SPDF, IRSA, NED and NEX[1]. The BDTF had some concern on each project's approach to preparing for the future so each program was asked to describe their approach. In particular, they were asked to respond to three specific questions:

1. What are the processes for planning for future (5-10 years) capabilities of your service? How and from whom do you gather input for this planning process and where does input typically come from? What new feature(s) do you really want to implement?
2. What feature(s) of your service would you like to stop performing? How do you gather input for making such decisions and where does input typically come from? What is preventing you from stopping?
3. What steps you are taking to make your data interoperable with allied data sets from other data sites in and out of NASA? How do you find allied data sets and what criteria make data sets candidates for enabling interoperability?

Much of the following was derived from the archive projects' responses to these questions as well as from the presentations by and interviews with the key scientists and managers associated with these projects.

Discussion

There is a wide variety of approaches for distributing NASA science data amongst SMD's four science disciplines. These approaches range from multi-mission, centralized storage managed directly by NASA to federations of data providers hosted by other institutions but managed by NASA.

Regardless of this diversity, most data archive programs rely on periodic peer reviews (such as senior reviews) to help shape long range planning when it comes to preparation for the future. In addition, there is mixed use of user surveys, help desks, user committees and feedback from direct interactions at major conferences to help guide each data center's plan. The strategy for each center appears to be appropriate for the size of their user communities which range from a

[1] ESDIS - Earth Science Data and Information System (https://earthdata.nasa.gov/about/esdis-project),
HEASARC - High Energy Astrophysics Science Archive Research Center (https://heasarc.gsfc.nasa.gov),
MAST - Mikulski Archive for Space Telescopes (https://archive.stsci.edu),
PDS – Planetary Data System (https://pds.nasa.gov),
SDAC - Solar Data Analysis Center (https://umbra.nascom.nasa.gov/index.html/),
SPDF – Space Physics Data Facility (https://spdf.gsfc.nasa.gov),
IRSA, NED and NEX - Infrared Science Archive, NASA Extragalactic Database, NASA Exoplanet Archive (https://www.ipac.caltech.edu)

few hundred specialists to millions of more general users. As the size of the community grows, there is clearly an increase in rigor to incorporate user feedback: smaller operations use informal approaches for collecting and incorporating feedback e.g. requests made at an AGU poster session that is incorporated when the developer has some free time, while larger programs may have designated user advisory committees making recommendations that are tracked and implemented under version control.

There is a general desire to expand the use of Python and machine learning in the data centers and in the wider community and to build up skills in data science, cloud computing and data security. The level of effort dedicated to investigating these new technologies varied, again mostly in line with the range in the size of each operation. Cloud computing efforts range from the purely conceptual ideas to "use the cloud", prototype demonstrations to beta-testing initial implementations such as what is currently happening in Earth Sciences. The BDTF also sensed that this is a rapidly changing situation that could move from concept to implementation on relatively short time scales.

Most data centers shared common problems of managing older mission data and maintaining old interfaces to archived missions. They also desire or are being forced to off-load services that could be better managed elsewhere and show an interest in "getting out of the business" of managing data stores and in avoiding "vendor lock-in." One example is the upcoming NISAR mission, which is expected to generate over 400TB/day. These issues can hold back the centers from their overall goal of modernize operations.

The data hosted by most data centers are fully integrated into the interoperable data environments of their respective science themes. This is achieved by a variety of methods, such as leveraging interfaces to virtual observatories and participation in the Open Geospatial consortium. The degree to which data discovery and appropriate data usage are supported varies among the disciplines. The Earth and planetary sciences share the problem of a wide variety of data types and areal coverage -- from in situ air- or spacecraft measurements, radar and lidar scans to global remote sensing -- that make interoperability challenging. Both directorates have evolved to a distributed system to manage this variety but at the risk of complicating discovery. The variety problem recedes as one pulls away from the planets, first to the mix of in-situ and remote sensing instruments in Heliophysics, and ultimately to purely remote imaging and spectroscopy in Astrophysics. The simpler underlying data structure in the latter disciplines aids the development of uniform discovery tools that fits well within a virtual observatory construct. However, the tools for discovery are still best within sub-disciplines of most directorates rather than across them.

The BDTF notes that the number of science papers, often from outside the original science team, that rely on archive data is increasing across the board, and is rivaling papers based solely on new mission data in astrophysics and heliophysics. This phenomenon, coupled with the other trends mentioned above (challenges with discovery and use of disparate data, difficulty leveraging older/archived data, and an increased need to apply common computational tools and platforms such as Jupyter Notebooks against data from different silos), highlights the need for seamless access between data archives and across multiple missions. There are examples in the Earth sciences, notably for climate modeling, of successful applications of cross-silo interoperability. To enable increased instances of cross-silo interoperability, NASA needs to encourage the use of uniform, human and computationally readable, data and metadata description standards and protocols for automated access to the SMD data archives to enable large-scale data studies. This becomes critical as the community moves to big data approaches. There is little hope of applying big data methodologies without some means of delivering uniform, reliable data. Any such approach should have clear goals for inter-silo interoperability and be mindful of cost, especially if enabling interoperability with much older archives is a cost driver.

Finding

The BDTF finds that SMD's data archive programs and projects are performing quite well and are properly taking steps to modernize and meet future challenges. Like all infrastructure projects they could use increased budgets and will put them to use wisely. Selective implementations of recommendations coming from peer reviews and the communities served are important to consider when augmenting what are generally steady-state funding profiles.

Finding

The BDTF finds that the fraction of science papers, often from outside the original science team, that rely on archive data is increasing in all divisions, and is rivaling the fraction of papers based solely on new mission data and in many cases, exceed 50%.

Recommendation

Given the empirical evidence that community use of these archives leads increasingly to discovering new science and combined with the aggregate budgets, the BDTF recommends that as a policy SMD should view the set of its data archives at the same rank as the flight missions in its portfolio.

Rationale

NASA data archives and centers are playing an increasing role in generating new scientific results as tools come available to discover and extract new meaning from existing data sets. The combined budgets of the data archives are comparable to that of mid-level missions during implementation! Promoting their status relative to new missions will provide to senior management both visibility and the focusing of appropriate resources for optimizing NASA science programs.

Consequences of not adopting recommendation

Valuable knowledge buried within science data archives may be overlooked or only rediscovered by new missions at great expense.