National Aeronautics and Space Administration

EXPLORE SCIENCE

NASA Archives Processing and Data Exploitation

Summary Report

August 2018

www.nasa.gov

All studies, findings and recommendations in these deliverables have been submitted to the Strategic Data Management Working Group (SDMWG) and to officials at NASA HQ. The opinions expressed in these materials do not reflect NASA's concurrence, approval, or indicate steps to implementation.

Background

Over the course of two days in August, the Strategic Data Management Working Group (SMDWG) within NASA's Science Mission Directorate (SMD) brought together members of the science community to discuss archives processing, data management and new, innovative initiatives on the cutting edge in science. The workshop's primary goal was to bring thought-leaders from Earth Science, Astrophysics, Heliophysics, Planetary Science, and High End Computing at NASA together to discuss both common and particular data management challenges each division faces to help shape the development of the Strategic Plan for Scientific Data and Computing at NASA's Glenn Research Center (GRC).

Summary

The two-day NASA SMD Archives Processing and Data Exploitation Meeting incorporated a variety of panels and individual presentations by each division within SMD. Ellen Gersten, the SMD Executive Officer, began the meeting by highlighting the four major objectives that will guide the 5-year strategy for all SMD data and computing:

- Continued free and open access to scientific data
- Improved ease of use and discoverability
- Enhanced science applications and new use cases
- Incorporating best practices and state of the art through partnerships

The break-out topics discussed throughout the day touched on the goals above and are listed in the *Session Themes* section. Representatives from each division covered topics, trends and challenges specific to their users, as well as current and future data needs. There was a concerted effort to make all sessions an active conversation on how to address challenges, meet future needs and standardize a data approach across SMD, all while highlighting inter-agency collaboration and coordination. The most frequently mentioned topics and concerns throughout the workshop sessions were as follows:

- Integration of commercial or NASA approved cloud
- Bringing analytics to the data
- Developing a standardized metadata approach across SMD
- The vast diversity of data across SMD
- How to maintain data usability
- Best practices and lessons learned from the Astrophysics Data System (ADS)
- Mandating a data management policy across SMD

To conclude the workshop, a post-meeting survey to gauge interest in discussed initiatives to identify and begin next steps (see *Appendix B*).

Session Themes

Division Overview and Division Deep Dive

The first two sessions in the workshop were devoted to, respectively, summarizing and describing in detail the challenges that SMD's divisions face from a computing and data perspective, and some of the solutions and approaches they are using to address those challenges. Divisions shared many commonalities, but still had significant differences in their makeup and needs - primarily, the amount of data being produced and how researchers are accessing it.

Each of the speakers highlighted the need to develop open-source cloud software for reuse across the Agency and by NASA-funded researchers. The speakers acknowledged the ongoing challenges with data formats, special time-scales, smoothing and moving heavy data around. They also discussed the challenge and necessity of engaging the data modeling community to develop tools that allow researchers to work directly with data. Finally, they emphasized the need for NASA to standardize their metadata structure through an established group or office.

Astrophysics has a digital library portal for researchers in Astronomy and Physics, ADS. ADS databases contain more than 14.0 million indexed and searchable records of publications. Many workshop attendees discussed the potential benefits of extending ADS to integrate other divisions of SMD (or even the entire Agency).

Earth Science has had some success in addressing data-related challenges in three of its major systems, the Earth Observing System Data and Information System (EOSDIS), the Earth Science Data and Information System (ESDIS) and Earthdata. EOSDIS provides end-to-end capabilities for managing NASA's Earth science data from various sources (e.g., satellites, aircraft, field measurements). Operations are performed within a distributed system of many interconnected nodes, Science Investigator-led Processing Systems (SIPS) and distributed, discipline-specific, Earth science Distributed Active Archive Centers (DAACs). The DAACs serve a large and diverse user community (as indicated by EOSDIS performance metrics) by providing capabilities to search and access science data products and specialized services. Best practices include cloud computing using the Earth Data Cloud, incorporating user feedback into design, the use of the metadata clearinghouse, Common Metadata Repository (CMR) as a requirement to export data, implementing a user registration system to track users incorporating digital object identifiers (DOIs) and citations. Earth science also issues surveys to learn more about the type of users accessing the system and what types of data they access.

Planetary Sciences offers several best practices that can be adopted by others across SMD, including a mandatory peer reviewed process, annual training that is available on-demand and planning for data archiving in the early phases of mission planning. The speaker from Planetary Sciences also suggested forming a data and archive working

group, writing an archive plan and defining data products to be archived and scheduled for public release.

Heliophysics archival data is divided by sub-disciplines based on different physical areas of the sun. In other respects Heliophysics' data archives are ad-hoc responses to missions, and even though they are a low budget priority, they often end up accounting for 40% of a mission's cost due to the longevity of the data. Challenges around archiving include diverse data formats, lack of metadata standards or a metadata bridge for interdisciplinary work, general metadata production and use and the difficulty of mining large data sets. Suggested best practices include senior reviews to assess whether a mission extension should be done, incorporating High-End computing resources as part of the senior review, established standards for data and metadata formats and more consistent implementation of Heliophysics' data policy.

The speaker from NASA's High-End Computing Center (HECC) noted that NASA maintains two copies of all its data (as a best practice and for disaster recovery purposes). The Agency hasn't deleted or destroyed any of the data since it was created. Much of the historic data has no metadata other than filenames, making it highly difficult to find and use. He proposes the best-case scenario is to copy datasets next to large-scale compute to meet the challenges and encourage more direct analysis capabilities (analyze in place). Because archive systems sit on high-performance networks with direct analysis capabilities, moving the data to the cloud will considerably increase productivity.

Citizen Science

During the Working Lunch, Marc Kuchner of SMD's Science Engagement and Partnerships spoke about the Citizen Science project and how it is essential to help users from all scientific backgrounds (i.e., novices, experts and super-users) properly use the NASA science data. Kuchner highlighted several opportunities across SMD for public involvement in NASA's science and technology programs, such as the Sun Spotter and Disk Detective projects. Citizen scientists use the data archives directly and write their own code for accessing NASA data. The challenge is that the projects produce more data than what needs to be archived. Kuchner recommends that the divisions seek funding and start planning a Citizen Science project before the data starts arriving to prevent being overwhelmed with responses and questions. The workshop attendees acknowledged the need for further discussion on developing user-interface tools and well-designed questions for citizen scientists who aren't trained in data science.

Data Management and Stewardship

The session examined how data stewardship can enable science, and how different types of science are impacted by data stewardship due to architectures, standards and requirements. Without data policies and data management plans, the divisions experience several problems such as long tail legacy data; making historical data interoperable; standardizing complex and high-maintenance share services; data storage and preservation instability; redundancy; cybersecurity; and loss prevention.

Recommended solutions included: encouraging management plans in proposals, defining the value proposition for science data management; increased focus on data preservation; as well as the sharing of code, data and data products. All speakers agree retaining, rebuilding and maintaining the data is important and will ultimately free up time and resources for scientific pursuits.

Participants agreed that to start addressing some of the issues around data management, SMD will have to define “Big Data”; figure out how to build data-sharing connections between clouds belonging to its academic, industry and international partners; develop a strategy for leveraging commercial data management solutions; solicit information on best practices from NASA partners; and start designing mission data collection with long-term data retention and future use in mind.

The Earth Science system EOSDIS employs a scaled agile framework and shared timeline and tracking tools, which help it keep track of metrics. The speaker from Earth Science emphasized the importance of content management structure with metadata and taxonomy, open APIs, common tools and cross-system Enterprise Services as best practices that all SMD divisions should implement.

Planning for the Future: User Access and Scientific Exploitation

The purpose of this session was to address the emerging trends in “Big Data,” data management and processing technology and how data can be harnessed to optimize scientific research. Each division discussed their uses of emerging technology trends. After making presentations, the speakers participated in a panel discussion moderated by Andrew Bingham (Jet Propulsion Lab) on the commonalities related to user access and scientific exploitation and identifying cross-division opportunities for collaborations. Knowledge sharing was a driver in each speaker’s presentation, and they highlighted how “Big Data,” multi-scale, multiple data sources and cloud computing are and will continue to affect NASA’s data and users’ needs. As data analysis needs increase, there may be more opportunities to practice interdisciplinary research. Centralizing data management by implementing ADS across SMD would make gathering metrics easier. The speakers agreed that not all divisions are the same, but when and if each division comes into the “Big Data” world, having such conversations will afford them the knowledge of how to sustain.

Incorporating emerging trends and best practices will support computation in data management, optimize data while maintaining provenance and synchronization and embrace machine learning as a pathway forward. New techniques including (but not limited to) machine learning will have a significant impact on data enhancement, distribution and archiving.

Partnerships (Discussion)

The attendees broke out into group discussions at each table to discuss synthesizing all the information covered over the two-day period from table discussions and panel

sessions. Attendees offered numerous best practices, success stories, challenges and potential pathways to move toward. In addition to summarizing findings, attendees also discussed partnership opportunities amongst NASA's divisions, other government agencies, colleges and universities and commercial companies. The overall findings and next steps are listed as follows and are in no specific order:

Overall Best Practices

- High End Computing Data Management Plan (DMP)
- Data Management Plans (DMPs)
- Publication Plans
- Digital Object Identifiers (DOIs)
- Discovery and Use
- Data dictionary
- Benchmark universities using cloud computing
- NASA investing in future data storage technologies

Overall Challenges

- Lack of budget dedicated to archives
- Data volume trend is increasing
- Open sources proposes security issues
- Lack of easy-to-use visualization tools
- Variety and complexity of the data
- Sustained funding for tools
- Data policy and metadata standards are needed across SMD
- Server-side analysis and machine learning on “Big Data”
- Seamless integrated access
- Cross-discipline science
- Rapid changes in technology
- NASA needs a Clearinghouse for pricing
- Lack of opportunities for brainstorming and networking amongst scientists and users
- Technical people need to be included and utilized in archival and data management conversations as they have the expertise

Success Stories

- Standard data formats
- Put requirements on DMPs for principal investigators (PIs)
- Document SMD-wide lesson learns for knowledge sharing
- White Papers to document successes and best practices around data management

Appendix A: Wrap Up Session

Overview

The two-day workshop wrapped up in a table group discussion format. Attendees were given the opportunity to synthesize and discuss suggested next steps for SMD.

Overarching Themes for Future Areas of Interest

Each table shared and submitted feedback on provided Post-It Notes. The feedback was bucketed into overarching thematic areas and recorded. The subsections below correspond with those themes, and the list of items in each subsection are feedback notes associated with those themes.

Interoperability

- NASA investment in data service research needs to be coordinated with the mindset of eventually providing an operational service to researchers
- Sharing results in the cloud ensures the authenticity of data as its accessed and curated by researchers
- The Heliophysics and Earth Science divisions could benefit from increased interoperability across the communities
- Increase diversity of platforms, sensors and data by leveraging solutions from the commercial sector (e.g., smallsats and cubesats)
- Accommodate interdisciplinary team approaches to data access and analysis
- Identify and implement standard access protocols to Distributed Active Archive Centers (DAACs) and the cloud to facilitate bringing compute and supercomputing to the data

Sharing Expertise

- Monthly data archive telecom
- Take turns sharing new developments, challenges, ideas, etc. across the divisions and NASA centers
- NASA IT Security and Procurement sharing would assist in centralizing knowledge and processes
- Leverage the inherent talent of this group to devise mission-setting solutions/options for the future
- Accept the notion that there is no “one size fits all” solutions to data
- There is a need for continued forums for SMD data centers to cross-fertilize and share ideas and best practices
- Synergy in understanding data stewardship, data publication, curation focus and priorities (user experience)
- Establish NASA center-based working groups to share ideas on data exploitation, archiving and cloud computing

- Convene technical cross-discipline meetings to discuss the data/metadata intersections between disciplines
- All proposed data policies should be reviewed for technical feasibility before they become mandates
- Need a mechanism to share knowledge, lessons learned and expertise across divisions
- Create a NASA data archive listserv or similar forum for NASA employees and researchers
- Host topical technical workshops on common cloud tools

Computing Cost Models and Concerns

- Explore vehicles for obtaining enterprise cloud computing licenses
- Agreements (Amazon Web Services, Google, Microsoft, etc.) to understand costs and projects
- Cost fluxes/changes
- Cloud computing cost estimation tools
- Address barriers to using commercial cloud across SMD
- Who will provide support to customers of NASA-provisioned cloud computing services?
- What is the process for budgeting in set spending cycles?
- How do calculate costs for cloud service use by with non-NASA partners?
- Overhead rates are different for services versus capital equipment (i.e., traditional in-house IT model)
- Share cloud cost model for groups to use periodically to evaluate costs for data storage, egress and computing
- Early High-End Computing Capability (HECC) engagement in mission processing cost modeling (Phase A or B)

SMD-Wide Data Management Policy

- Data and algorithms
- Data management policy must also consider theory, lab results, models, model runs
- SMD-wide peer review for all new missions
- High-level SMD data policy
- SMD-wide data policy (uniform level of service for data)

Leverage High-End Computing Capability (HECC)

- Explore opportunities to take advantage of HECC
- Clarification of specific infrastructure, services and costs
- HECC as potential long-term data storage and backup

Future Meeting General Topics and Suggestions

- Allow more time for Q/A and breaks at future meetings

- Discuss implementation of science platform (computing near the data)
- Hold more regularly scheduled SMD data workshops
- More discussion of what “Big Data” is and how should we address it
- Time variant data can assist in moving toward standardize metadata
- Extract lessons learned from Earth Science as a benchmark
- Need to discuss the various architectures for cloud-use presented in the workshop
- Explore hybrid cloud, multi-cloud
- How to evaluate metrics
- Open science grid concept applied to archives and computing

Machine Learning (ML)

- Expert system integrative learning (active learning) with teams
- Expert pre-created data training sets, cleaned and normalized data for ML, system-wide data fusion

Maintain Current Level of Service

- Make sure NASA’s current requirements and goals are adequately addressed in policies
- Need to expand into new capabilities while maintaining overlap with old/legacy systems
- Address resource and money needs
- How to enable science analysis across large datasets (NASA and others)
- Develop environments for rapid prototyping to support and attract young scientists who want agility

Common NASA Cloud on Ramp

- What are the success criteria for hosting data in the cloud?
- What is the concept for a long-term archive?
- Partner to understand needs and develop user-focused data analysis and research software
- Develop set of training/tools (or point to existing) to help users and builders with data-intensive functions
- NASA should be running a research cloud for its programs to mitigate price structure issues

Common Metrics

- Consolidate customer satisfaction surveys
- Common software for user metrics

Industry Engagement

- Look to NASA’s industry and academic partners for cloud success stories

- Approach for engagement with multiple cloud venders to get help in developing data analysis tools that users need
- Look for lessons learned

Astrophysics Data System (ADS)

- DOIs work for research papers
- User metrics
- SMD-level ADS – modify the system to be available all 4 divisions (Aware Earth will be the hardest to add)
- Open source collaboration
- API to support other clients
- Develop unified (Earth Science, Astronomy and Planetary discipline) thesaurus
- Need to gather information on who uses our data and the impact
- ADS impact metric
- Need SMD-wide approach to research citations for data usage and efficiency of data management metrics
- 25% of ADS data citations come from Moderate Resolution Imaging Spectroradiometer (MODIS) data
- Closed Native projects
- Cross-division data sharing is good but cross-Agency sharing is critical
- Science use cases should be cross-divisional or cross-Agency
- Exoplanet research on cross-cutting challenges
- Cross-cutting opportunity to work with National Science Foundation Origin of Life grant program
- Opportunity for cross-divisional ground-breaking science

Appendix B: Post-Meeting Survey

At the end of the two-day meeting, SMD leadership charged the attendees to continue the week's conversations after the meeting and gathered their emails. A short survey was developed to collect information to help guide the SMDWG's actions and recommendations on follow-up activities and next steps. The survey was hosted on Google Forms and was broken up into two sections. The first section tailored questions to assess the attendees' interest in continued engagement with specific working groups. The second section asked for suggestions on how to make future gatherings of the SMDWG more effective. Of the 80 attendees, 44 submitted responses.

The breakdown of responses is as follows:

- 14 agreed to help lead at least one of the three working groups
- 7 agreed to help with the data Architectures and Cloud
- 6 agreed to help with the ADS working group

Some common requests for changes in future workshops included:

- Focus more on specific topics (and possibly break participants into smaller groups)
- Divide sessions by common activities, technologies, challenges – not by SMD divisions
- Make these workshops regularly scheduled (not one-off) events

Appendix C: Speakers (Listed by Session Grouping):

- Ellen Gersten, Science Mission Directorate
- Pat Knezek, Astrophysics Division
- Kevin Murphy, Earth Science Division
- Bill Knopf, Planetary Science Division
- Jeffrey Hayes, Heliophysics Division
- Tsengdar Lee, High End Computing
- Hashima Hasan, Astrophysics Division
- Andrew Mitchell, Earth Science Division
- Ray Walker, Planetary Science Division
- Piyush Mehrotra, High End Computing
- Marc Kuchner, Science Engagement and Partnerships
- Tom McGlynn, Astrophysics Division
- Katie Baynes, Earth Science Division
- Bob McGuire, Heliophysics Division
- Bob Ciotti, High End Computing
- Ellen Salmon, High End Computing
- Chris Lynnes, Earth Science Division
- Jack Ireland, Heliophysics Division
- Harry Teplitz, Astrophysics Division
- Josh Peek, Astrophysics Division
- Steven Hughes, Planetary Science Division
- Dan Duffy, High End Computing
- Rahul Ramachandran, Earth Science Division
- Andrew Bingham, Earth Science Division


NASA
EXPLORE
with us