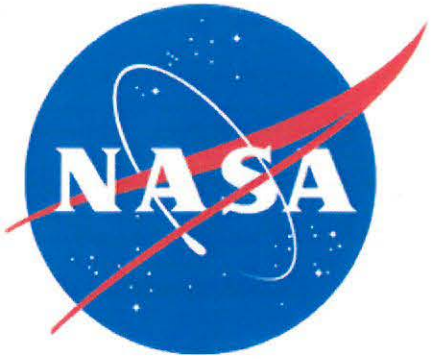

Science Mission Directorate's

Strategy for Data Management and Computing for Groundbreaking Science 2019-2024

Prepared by the Strategic Data Management Working Group

Approved by:

12/17/19

Thomas H. Zurbuchen, Ph.D.
Associate Administrator,
Science Mission Directorate

Table of Contents

INTRODUCTION

Background

NASA's Science Mission Directorate (SMD) seeks to expand human knowledge through new scientific discoveries in order to understand the Sun, Earth, Solar System, and Universe. SMD, in partnership with the Nation's science community, conducts scientific studies of the Earth and Sun from space, returns data and samples from other bodies in the solar system, and peers out into the vast reaches of the universe. This work seeks to address three core contexts that span the breadth of our activities:

- Discover the secrets of the universe
- Search for life elsewhere
- Protect and improve life on Earth

SMD's missions and research activities inspire curiosity and increase the understanding of our planet, the solar system and universe. A core capability of SMD is the ability to collect, store, manage, analyze and distribute data and information for scientists, international partners and industry to further science and increase knowledge. Each of the four science divisions within SMD generate, analyze, and archive large amounts of data to support unique science objectives and delivers data and scientific results to millions of users around the world. As the NASA Advisory Council's Ad-Hoc Big Data Task Force observed[1], the fraction of science papers that rely on archive data is increasing and, in many cases, exceeds the fraction of papers based on new mission data. That is, half of all science results are coming from archive data.

SMD currently stores over 100 Petabytes (PB) of observational data and model results. Based on projections of data rates for new missions in development, within 5 years, all four science divisions are cumulatively projected to generate over 100 PB of data per year and continue to grow rapidly as additional missions are launched and new models are run. This anticipated growth of NASA's science archives presents unique opportunities for new scientific discovery as well as significant challenges for data management, curation, access, analysis, maintenance of provenance and computing.

SMD also supports supercomputing facilities for research needs. Two major supercomputing centers provide more than 10 Petaflops aggregated peak capacity for SMD's scientific research and engineering workloads. These centers work mainly in scientific research areas of astrophysics and heliophysics theory development and validation, weather and climate modeling and data assimilations, and large-scale data synthesis and analysis.

[1] https://smd-prod.s3.amazonaws.com/science-red/s3fs-public/atoms/files/6th%20and%20Final%20BDTF%20report%20to%20SciComm%20171128%3DTAGGED.pdf

Overview

Over the past decade terms such as data science, machine learning, and cloud computing have become widespread. The technologies and practices underlying them are driven by significant commercial investment and a recognition that data and computing capabilities are capable of sparking curiosity and innovation not imagined previously. Along with these technological advancements, expectation of the scientists, commercial and scientifically interested users of public data have changed. There is a recognition that publicly funded research and the underlying data should be open and easily accessible by larger communities of users to support innovation. Conducting science in the open builds trust, advances science and allows access to public information for academic, international and commercial partners. The pace of technological and social change will continue to accelerate as new technologies are developed by NASA and externally.

The Science Mission Directorate (SMD) has long recognized the importance of its data resources and now seeks to take a more strategic view of its science data systems, including high-end computing, to promote more efficient and effective data management across the four science divisions, as well as to enable cross-disciplinary discovery and analysis of science data, including the use of High-End Computing. To aid in this work, in January 2015, the Big Data Task Force (BDTF) was chartered through the NASA Advisory Council (NAC) to study and identify best practices in big data. The final report of the BDTF was delivered to the NAC Science Committee in November 2017.

Strategic data management and science computing across SMD also was identified as a priority for assessment and action during the senior leadership retreat in May 2017. In February 2018, SMD chartered the Strategic Data Management Working Group (the Working Group) with representatives from each division to develop a new directorate-wide strategy to enable greater scientific discovery over the next five years by leveraging advances in information technology to improve NASA's science computing and data archives.

This document represents the strategic recommendations of the Working Group. SMD's *Strategy for Data Management and Computing for Groundbreaking Science* complements NASA's Vision, "To discover and expand knowledge for the benefit of humanity," by creating a resilient foundation for SMD data, information and computational capabilities through open data and software policies, continuous evolution, partnerships and commitment to users. Implementation of this strategy will align advances in information technology with the unique needs of science data systems and computing, and is designed to inform future investment strategies to enable greater scientific discovery.

Approach

Development of this strategy was guided by four principles:

- Improve discovery and access for all SMD data to immediately benefit science data users and improve the overall user experience

- Leverage current technology for the discovery, access, and effective use of NASA's data, as well as enable new technology and analysis techniques for scientific discovery
- Identify large-scale and cross-disciplinary/division science users and use cases to inform future science data system capabilities
- Champion robust theory programs that are firmly based on NASA's observations

Given the breadth and potential impacts across the scientific community as a result of this strategy, the team used several mechanisms to collect stakeholder feedback and to promote data sharing and information gathering:

- Archives Processing and Data Exploitation Meeting, August 9-11, 2018
- NASA Advisory Council Science Committee draft analysis and findings in response to the report of the BDTF, August 28, 2018
- National Academies of Science, Engineering, and Medicine's *Open Source Software Policy Options for NASA Earth and Space Sciences*, September 25, 2018
- Workshop on Maximizing the Scientific Return of NASA Data, October 30-31, 2018
- Request for Information (RFI): Strategic Plan for Scientific Data and Computing, September 18-November 1, 2018

Through this process, over 450 people from academia, U.S. and other government agencies, commercial providers, professional societies, and the general public shared their ideas to inform development of the plan. These people represented all four science divisions, as well as the High-End Computing, information technology, and data science communities. Summaries from the NASA workshops and the RFI responses are available at:
https://science.nasa.gov/researchers/science-data

CURRENT STATE

Data and computing systems supporting the Science Mission Directorate (SMD) have traditionally embraced a systems tailored for each science division. These uncoupled systems have allowed each science division flexibility to support unique science requirements and user needs of the communities. Management of data and computing resources is typically based on the specific needs of each mission or division, with limited consideration for enabling inter-divisional research. When compared to peer agencies such as the National Science Foundation and the National Institutes of Health, NASA's investments have traditionally been focused on missions rather than cyberinfrastructure and NASA does not have equivalent strategic-level programs to invest in the adoption of cutting-edge technologies (*i.e.*, deep learning, machine learning, and artificial intelligence, and in applying these techniques in a High-End Computing environment).

In Fiscal Year 2019, the Science Mission Directorate invested $120.7 M for its computing and data management activities across the four science divisions, as well as $69 M for High-End Computing. The Working Group conducted a current-state assessment to understand how each division manages its computing and data resources. Interviews with representatives from all four

Science Divisions were conducted to investigate their data usage policies and common practices. Tables 1 and 2 summarize the key findings from this assessment.

Table 1. Current State Assessment - Metrics

	FY19 Budget ($M)	# of Data Centers/Nodes	# of Archival Data Programs	# of Users (M/year)	2019 Size of Archive (TB)	2019 Annual Ingest Rate (TB)	2024 Projected Volume (TB)
Astrophysics	25.2	4 Archives and 4 Data Publication Services	3 – ADAP, HST, Chandra	10.56 (includes ADS)	7000	4 – 150, per archive	30000
Earth Science	70	12 DAACs	1 – EOSDIS	7.2	27400	8285	250000
Heliophysics	6.5	2 Active Archives	2 – HDEE, GI	4.8	19000	4000	44000
Planetary Science	19	7	6 – PDART, MDAP. DDAP. RDAP, LDAP, and NFDAP	3.21	1600	250	30

Table 2. Current State Assessment - Policy

	Data Access Approach	Data Management Approach	Division-Level Standardized Data Management Plan	Cloud Computing Approach	Publication Discovery Status	HEC Demand[2]
Astrophysics	Online through APD data archives and mission websites	Jointly managed by Program Scientist, Deputy Program Scientist & Program Executive	No	Ongoing projects by individual APD archives	Accessible via ADS	High
Earth Science	Online through EarthData Search	Management duties are performed by ESDIS	Yes	Migration of high value data sets is underway. Central to new mission.	Working to improve access for new publications via digital object identifiers	High
Heliophysics	Online through NASA and mission websites; offline through direct requests to PIs	Management approaches are directed by each mission	Yes	Currently supporting use cases at the division level	Hard to discover; not tied to missions or data; journal access varies	High
Planetary Science	Online through science node sites and the PDS site	Distributed management senior review in the next year	No	Currently investigating use	Use of digital object identifiers to cite PDS data; PDS does not track publications	High

[2] All four divisions have significant back log of computational jobs and the available computing resource cannot meet the demands. The expansion factor ((wait time plus run time)/run time) is usually greater than 2.

VISION AND GOALS

The Working Group has developed a vision, mission, three goals, and 11 associated strategies for strategic data management and computing across all of the Science Mission Directorate (SMD). Additional tactical guidance necessary to implement the proposed strategies, as well as the current state of those efforts, can be found in the Findings and Recommendations section of this document.

Vision

To enable transformational open science through continuous evolution of science data and computing systems for NASA's Science Mission Directorate.

Mission

Lead an innovative and sustainable program supporting NASA's unique science missions with academic, international and commercial partners to enable groundbreaking discoveries with open science data. Continually evolve systems to ensure they are usable and support the latest analysis techniques while protecting scientific integrity.

Goals

We endeavor to meet the vision and mission through collaboration and responsible investments in technology – recognizing we are the stewards of irreplaceable data, unique knowledge, taxpayer dollars and scientific integrity. Based on guidance from science communities, the National Academies, academia and commercial and international partners we will advance the management and computation capabilities to support groundbreaking science at NASA for the benefit of humanity in three ways:

1. Develop and Implement Capabilities to Enable Open Science
2. Continuously Evolve Data and Computing Systems
3. Harness the Scientific Community and Strategic Partnerships for Innovation

Goal 1: Develop and Implement Capabilities to Enable Open Science

NASA is required to make scientific data and software open, as directed by the Administration, Congress, and as recommended by the National Academies of Science, Engineering, and Medicine. The National Aeronautics and Space Act of 1958 specifically directs NASA to provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof. Making science data, software and information discoverable, open and accessible encourages collaboration and innovation, as well as increases transparency.

Significant work is needed to develop systems that provide the foundational components of an SMD open science ecosystem. At the heart of the ecosystem are modular open services that can

be used individually or combined to support improved user experience for the discovery, access and analysis of data and information. Some of these components already exist within SMD but are isolated to specific project or divisions. Investments are needed to make the existing systems more robust to support cross-divisional requirements and to develop capabilities that are missing. There will still be a need for unique specialized capabilities for missions, however by utilizing a modular open services architecture, the unique and generalized requirements can be addressed. Generalized systems described in the sub-goals will be used to support higher level NASA and U.S. Government requirements for open data and software in a consistent and scientifically accurate manner. This transition will take time and represents an investment in capabilities that will support cross-domain discovery, use and analysis of SMD science data for science, commercial and any other use.

Strategy 1.1: Develop and implement a consistent open data and software policy tailored for SMD building on Agency (NPD 2230.1) and U.S. Government guidance (e.g. OMB-M-13-13, Making Open and Machine Readable the New Default for Government Information, P.L. 115-435, Evidence Based Policy Act Article II). This policy should be informed by the National Academies of Science, Engineering, and Medicine's 2018 report *Open Source Software Policy Options for NASA Earth and Space Sciences*. This SMD-tailored policy is the framework for the open science ecosystem and will be used to inform science data and software policies for system development. Specific implementation details and exceptions are outside the scope of this document.

Strategy 1.2: Upgrade capabilities at existing archives to support machine readable data access using open formats and data services in compliance with the Evidence Based Policy Act Article II. These capabilities will be tailored to the unique scientific requirements of each division to maintain scientific integrity while supporting widely accepted standards needed for access. Data and services will have sufficient functionality and performance to support the access methods, tools and techniques and be coordinated among Divisions.

Strategy 1.3: Develop and implement a SMD data catalog to support discovery and access to complex scientific data across Divisions. This catalog will provide consistent discovery methods for distributed systems utilizing standards, be machine readable and support fast, discovery of data archived by all SMD Divisions. A consistent SMD data catalog will support accurate representation of SMD data holdings in higher level catalogs needed for compliance with the Evidence Based Policy Act, including http://data.gov while simultaneously allowing for specialization needed for discovery of and access to data that spans SMD. These capabilities will be based on existing standards where feasible and development of new standards in collaboration with the appropriate communities when needed.

Strategy 1.4: Increase transparency into how science data are being used through a free and open unified journal server. Such a system enables the public to freely access journal articles based on NASA data. Open access provides the public with clear evidence of the linkages between mission investments and scientific results. Broad access also facilitates cross-disciplinary research by removing barriers to collaboration. Recognizing that there are limitations to how many articles a researcher can reasonably be expected to read in a year, migration to a single server that employs machine-assisted learning will also benefit the community. The computing

industry and library science have already advanced to a reasonably mature state, so NASA should proactively research and develop these technologies to organize information and knowledge in an easily searchable way.

Goal 2: Continuously Evolve Data and Computational Systems

Once operating, missions frequently collect irreplaceable data well beyond prime operations into extended mission phases (years to multiple decades). Therefore, data are used by a diverse community whose expectations can change significantly over time as the academic, commercial, and international sectors develop innovative, new data and computing capabilities.

Given the long duration of many science missions, the rate of innovation, and the opportunity to advance science with new tools, SMD must encourage the adoption of new techniques for existing data while balancing the need to maintain the scientific integrity of irreplaceable data. This will require near-term changes for new missions to continuously evaluate data and computational systems throughout the mission lifecycles to identify and use new techniques and technologies that are beneficial. Such evolution can be difficult incorporate into the ground data and computing systems for missions that have been in operations for many years unless the expectation for continuous evolution is considered from the beginning.

Innovations have the potential to have pervasive and radical impacts on the organizations that maintain, develop and operate and analyze data over long time periods. SMD must therefore continuously evolve data and computational systems to realize the potential of innovative techniques to more efficiently manage data and computing resources and establish policies optimized to support investments in technology development and adoption. This will require investments in data systems, computational approaches, and the workforce that harnesses technology are needed to support the evolution of data management and computing systems. Engaging with data scientists throughout this process is essential to ensuring that NASA Science is able to do this effectively.

Strategy 2.1: Establish standardized approaches for all new missions and sponsored research that encourage the adoption of advanced techniques. SMD missions should take a lifecycle approach in planning how their data will be acquired and managed for the duration of the project and after, with the consideration of their legacy in mind. Similarly, sponsored research should be supported to move in the same direction (*i.e.,* what is the legacy of the research and how can it be made accessible to everyone).

Strategy 2.2: Integrate investment decisions in High-End Computing with the strategic needs of the research communities using this capability. Technology investments should align with research needs that require HEC assets to maximize their potential return.

Strategy 2.3: Invest in capabilities to use commercial cloud environments for open science to make data accessible by diverse set of academic and commercial users.

Strategy 2.4: Provide tools and training to scientists to be better able to collaborate with all types of computational and computer scientists to enable the funding of successful collaborations, including Artificial Intelligence and Machine Learning (AI/ML).

Goal 3: Harness the Community and Strategic Partnerships for Innovation.

SMD and the individual science divisions do not operate in isolation and therefore should recognize there is tremendous value in engaging with multiple stakeholder groups to identify opportunities to increase collaboration and use of advanced tools and techniques to drive scientific discovery. The decisions on when and how to collaborate should be made in such a way that SMD sets policies and facilitates sharing best practices, while providing the science divisions with responsibility and flexibility to manage their systems to meet the needs of their communities. This will put management capabilities in the hands of those who best know the data, while allowing SMD to deal with exogenous risks and opportunities (*e.g.,* OCIO policies, NASEM recommendation, OMB guidance, software license, open software communities, training, and security) in a coordinated manner on behalf of the entire NASA science community. The community within NASA has deep understanding and understands the unique complexity of data and how to best use this information. By working across divisions, it will be possible to support the careful long-term stewardship of irreplaceable data.

Strategy 3.1: Cultivate a strong community of practice across SMD, the science archives, and the broader research community. As sections of the community come together, there is a growing realization that collaboration is the best way to tackle extremely large and interesting science questions. To enable increased collaboration, such a community of practice provides a forum to share best practices, discuss common strategies or concerns, and identify training needs.

Strategy 3.2: Partner with academic, commercial, governmental and international organizations to augment SMD's in-house capabilities. External sectors have expertise that is well-suited to managing, analyzing, and assimilating very large data sets and has expressed interest in working with NASA. These organizations often have unique expertise and complementary data that can support groundbreaking science.

Strategy 3.3: Promote opportunities for continuous learning through collaboration with academia, industry and other government agencies. Data analytics, computing, software development and data management are changing rapidly and accelerating the rate of scientific discovery beyond what any one individual can synthesize. Breakthrough science is now being done at the intersection of data science and traditional physical sciences.

FUTURE WORK

The Working Group recognizes that needs and best practices today will not necessarily be the same in five years. Therefore, periodic reassessment of the Science Mission Directorate's (SMD's) data and computing needs is warranted to ensure currency. The Working Group also encourages each science division to conduct similar reviews over time to evolve their capabilities.

Based on the results of our information gathering process, the Working Group did not identify immediate needs to support cross-disciplinary investigations. However, we believe that SMD's interest in enabling scientific discoveries at the boundaries between science disciplines will require the ability to integrate multiple datasets and SMD should incentivize community engagement in this regard. Migration to common metadata standards also provides the opportunity to consider developing a single data repository for all NASA science data, but this is a longer-term aspiration.

FINDINGS AND RECOMMENDATIONS

The Working Group provides 12 findings and associated recommendations for consideration in implementing the previously defined strategy. These findings and recommendations are in five areas: 1. Open Data/Open Software; 2. High-End Computing; 3. Archives Modernization; 4. Advanced Capabilities; and 5. Management.

In developing these recommendations, the Working Group identified three key operating principles that should be considered as part of the implementation process:

1. Do no harm to the data you have collected;
2. Take advantage of new missions to do evolution; and
3. Modernize historical data sets of high scientific value while protecting integrity of irreplaceable data.

AREA 1: Open Data/Open Software

Finding 1: NASA is subject to numerous executive orders and other policy direction requiring that data be open and accessible. Agency-level policy and direction is driven by the Office of the Chief Information Officer, but within NASA Science, the open data policies were developed independently and are not standardized across the science divisions. Additionally, new Federal guidance is currently in development for data and publications.

> **Recommendation 1a:** NASA Science should develop a standard open data policy for all new missions, and current operating missions also should also be encouraged to follow the policy. For new missions, the data requirements should be documented in the standard AO and SALMON templates, and compliance should be evaluated as part of PLRAs and appropriate lifecycle reviews. NPR 7120.5, and other NASA guiding documents, should be updated to include data systems in the compliance matrix.
>
> **Recommendation 1b:** Standard open data requirements should be included in ROSES for all new solicitations and subsequent award requirements. These requirements should reflect the most current guidance and consider the recommendation from the NASEM report.
>
> **Recommendation 1c:** Missions should take a lifecycle approach in planning how their data will be managed for long-term curation after KDP-F. The Earth Science Division's data management schedule provides a reference for how this can be done and should be used as a template as each division develops their own approach as is available at: https://earthdata.nasa.gov/collaborate/new-missions.

Finding 2: The broad science community is moving in the direction of open source and collaborative tools. For example, the DoD adopted open source principles[3] in 2008. The National Academies of Science, Engineering, and Medicine have called on NASA to adopt an open source policy. NASA is also subject to executive orders and other policy direction

[3] https://dodcio.defense.gov/Open-Source-Software-FAQ/

requiring that new code be open and accessible (e.g. August 8, 2016 Office of Management and Budget M-16-21 Memorandum).

Recommendation 2a: SMD should adopt a requirement for all new software development to be open source, except in instances of ITAR, EAR, national security, PII, or other similar restrictions. This is consistent with OMB-M-13-13. Any waiver to not having publicly accessible software will need to be approved by the Deputy Associate Administrator for Research.

Recommendation 2b: Any software developed under the ROSES NRA must be released as Open Source Software (OSS); opt-outs and alternatives must be justified and approved by the Deputy Associate Administrator for Research. The Earth Science Division currently has an OSS policy that could be adapted for a Directorate-wide policy[4]. This software policy must be added to the standard ROSES language. NPR 7120.8, and other NASA guiding documents, should be updated to include software in the compliance matrix.

Recommendation 2c: For new missions, an open source requirement must be included in the standard AO and SALMON language, and compliance should be evaluated as part of PLRAs and appropriate lifecycle reviews. For missions already in operation, software accessibility should be part of the senior review process. NPR 7120.5, as well as the PE Handbook and relevant mission documents, should be updated to include software in the compliance matrix.

AREA 2: High-End Computing (HEC) Program

Finding 3: High-end computing is a resource in great demand by SMD's research communities and SMD is one of the two major users of HEC resources. HEC serves a specific need for NASA Science's modelling and theory programs. HEC research in the nation often drives other new and advanced information technology research and developments. HEC is a member of the Agency's Capabilities Portfolios. Over the past decade, NASA's HEC program has been focused on buying hardware as opposed to integrating it into a research ecosystem. This is an unintended consequence from the establishment of the Shared Capability Assets Program within the agency. However, there is now a concerted effort in the Agency to rebuild some of the research and development efforts in the capability portfolios.

The current investment model is driven by Agency-level needs and is decoupled from the SMD research programs. This makes the true requirements for HEC resources difficult to manage. The last directorate-wide requirements assessment was done in 2013. Since that time, requirements for HEC resources and modern computational technology have evolved. For example, cloud computing was in its infancy 10 years ago.

Recommendation 3: SMD should conduct assessments no less than every five years for high-end computational resources. These assessments should evaluate HEC capacity, computing needs, and allocations across science divisions. This information should be used to develop a strategy for tracking and allocating cycle time for NASA resources and

[4] https://earthdata.nasa.gov/collaborate/open-data-services-and-software/esds-open-source-policy

exploring whether there is the demand for additional or new computational resources to support SMD's objectives. The output of this assessment will also be used as part of Agency-level planning through the Office of Strategic Infrastructure's Shared Capability Assets Program to determine if additional resources are required.

AREA 3: Archives Modernization

Finding 4: The majority of data from high quality research activities are already archived and made available to public. However, some data are not moved to archives and the scientific community requires additional incentivization to deposit data into these archives. In particular, historical data for most divisions has tended to be held exclusively by the original PIs. The community may also require additional funding to address some of the new archiving requirements to enable open data. Some ROSES elements already have sufficient funding to take care of high-level archiving of data, but this is not uniform across the entire portfolio.

Recommendation 4a: Digital data derived from NASA-funded research is required to go to a NASA archive for long-term curation and public availability. There are shared responsibilities between the principal investigators (PIs) and the archives to ensure transition of data. Archives must provide clear guidance on requirements to ingest data and investigators must adhere to the requirements. The Earth Science Division provides one example of such guidance at https://earthdata.nasa.gov/collaborate/new-missions. Additionally, the current archives must prepare to receive and validate data from heterogeneous sources, which may involve more "hands-on" work with PIs submitting data, metadata, and documentation. Such additional work will require additional funding, which should be addressed in the PPBE process.

Recommendation 4b: SMD should include an evaluation criterion for future ROSES solicitations to assess the adequacy of data management plans associated with individual proposals to support delivery of data to a final archive. SMD should evaluate whether an augmentation of funding may be required, either at the award or program element level, to enable this. As this becomes the norm, SMD should reassess the level of funding required.

Finding 5: Data analytics, computing, software development and data management are changing rapidly and accelerating the rate of scientific discovery beyond what any one individual can synthesize. This applies across the sciences, not just NASA-funded research. Given this rapidly evolving environment, we find ourselves in a situation where the current researcher-led model is not as sustainable. For example, the ways students are being trained in the sciences are not necessarily up-to-date given the evolution of data science. However, there is movement in that direction, especially by individuals in the research community, but there is not broad awareness of the overlaps.

Recommendation 5: SMD should strongly encourage collaboration and cooperation of data professionals in academia, industry, and elsewhere to enable cross-cutting scientific discovery. There are currently small-scale efforts in the scientific communities to provide data stewardship training, including cleaning and curation, but these will need to be expanded to support anticipated future demand. This activity should be a sustained to ensure

that new techniques (science or data), methods and algorithms for analysis, and management of data can be incorporated as they are developed.

Finding 6: There is an imbalance between the perceived value of developers and data stewards in relation to science researchers. This limits NASA's competitiveness and ability to attract the necessary talent to develop and support groundbreaking science. Given the role of the science data archives to enable new science, SMD is not fully utilizing the large investments already made in the data by not increasing the visibility and importance of this work.

> **Recommendation 6:** SMD should communicate the value of this capability and work with the NASA Centers to increase their recruitment of data science professionals. While NASA is at a competitive disadvantage for these types of positions, there are unique opportunities provided to data science professionals that only exist in the NASA environment.

Finding 7: The senior review process provides an opportunity to evaluate how well operating missions are generating data and performing operations. Issues related to data and software usability, discoverability, and documentation are often not considered as part of the current process.

> **Recommendation 7:** A process should be developed for data and software usability, discoverability, and accessibility to be evaluated as part of future senior reviews. This should also include utilization of advanced computing capabilities (e.g. high-end computing and commercial cloud computing). Information gathered through this process should be used to inform decision-making by the relevant archives.

Finding 8: Science data are generated on a mission-by-mission basis. Cross-mission data fusion projects enable more systems-level science that couples theory, modeling, and observations across disciplines, but they are the exception, not the rule. For example, in Earth Science, some systems-level science is being done, and there is movement towards more integrated, whole Earth system models and integrated Earth System analysis capabilities. However, while both Heliophysics and Earth Science study the Earth's mesosphere and ionosphere, rarely do the two collaborate because the data is not discoverable across both communities.

> **Recommendation 8:** In order to enable cross-disciplinary science, SMD must facilitate greater discoverability of like data holdings in various archives. SMD should engage with the science community to develop a metadata description that would cross science

Finding 9: SMD and the scientific community at large are interested in having better transparency into how science data are being used. Tracking of publications based on NASA data is inconsistent across individual flight projects and divisions. Further, access policies vary between publications, including differences in pricing structures and when articles become available for free. For example, some publications charge different amounts based on when an article was published while others allow free access to preprint copies of articles and charge for the final, printed version. While some communities already have systems that allow for open access to journals, this varies by discipline and subdiscipline.

Recommendation 9: SMD should create a free and open, unified journal server along the lines of PubSpace, ADS or ERS to make science papers more accessible to the public. NASA Science should also consider adopting the National Science Foundation's requirements to reinstate the need for grant recipients to provide copies of all published research as part of their annual reports.

AREA 4: Advanced Capabilities

Finding 10: The current grant funding structure does not encourage the use of heterogeneous or large datasets because the level of effort required to organize and prepare data for analysis is prohibitive for all but the most well-funded and sophisticated users. Cloud computing offers one opportunity to broadly improve access and analysis of very large data sets through server-side analysis removing the need to move and manage large data sets. Breakthrough science will come through application of new techniques, such as Artificial Intelligence and Machine Learning (AI/ML), to these datasets in highly scalable environments such as cloud computing and HEC. Pilot projects have shown that the use of AI/ML requires both specialized skills and that data be properly calibrated prior to analysis. This is a fundamental change to the way science has been done and attention needs to be given to the importance of credible and reproducible results.

Recommendation 10: SMD should encourage the science divisions to explore novel computational techniques, including cloud computing and AI/ML, through various avenues, including ROSES NASA Research Announcements, technology calls in cooperation with the Office of the Chief Information Officer and Space Technology Mission Directorate, and mission Announcements of Opportunity.

Finding 11: The advance of AI/ML has yet to be fully appreciated and understood by SMD and the science disciplines. The possibilities of serendipitous, cross-disciplinary science in such an environment is unexplored and needs to be understood.

Recommendation 11: SMD should make investments to incentivize and educate the community on how to use AI/ML to approach science in new ways. Hands-on training can be achieved through expansion of hackathons, competitions, and grant programs. Science results and lessons learned about the use of AI/ML will be shared at community meetings to increase awareness of the potential of these techniques.

AREA 5: Management

Finding 12: Each Science Division has been given autonomy to manage their data and High-End Computing (HEC) needs. This enables the Division to be responsive to the needs of their user communities, and they have been very successful. Nevertheless, the archiving and HEC communities representing each discipline do not regularly interact, which reduces insight into common opportunities and problems. There are opportunities for increased collaboration across divisions on similar data and HEC management activities. There is not, however, community support for the consolidation of these activities into a single organization within NASA Science to coordinate.

Recommendation 12: SMD should appoint a directorate-level Scientific Data and Information Officer to oversee this effort and serve as an interface between the Office of the Chief Information Officer, Office of the Chief Scientist, international partners, commercial providers, and others. The Scientific Data and Information Officer's responsibilities should include: 1. Setting policy for SMD, verifying compliance, and maintaining awareness of external policy drivers; 2. Maintaining a cross-divisional innovation forum to identify, select, and fund new opportunities; 3. Investing in targeted capabilities and tracking their progress; 4. Cultivating SMD's data and computing community through workshops, studies, training, etc., and; 5. Conducting periodic independent evaluation of the structure and content of the SMD data and computing portfolio.

APPENDIX A: Status of Findings and Recommendations

Status current as of December 17, 2019

AREA 1: Open Data/Open Software

Recommendation 1a: NASA Science should develop a standard open data policy for all new missions, and current operating missions also should also be encouraged to follow the policy. For new missions, the data requirements should be documented in the standard AO and SALMON templates, and compliance should be evaluated as part of PLRAs and appropriate lifecycle reviews. NPR 7120.5, and other NASA guiding documents, should be updated to include data systems in the compliance matrix.

Recommendation 1b: Standard open data requirements should be included in ROSES for all new solicitations and subsequent award requirements. These requirements should reflect the most current guidance and consider the recommendation from the NASEM report.

Recommendation 1c: Missions should take a lifecycle approach in planning how their data will be managed for long-term curation after KDP-F. The Earth Science Division's data management schedule provides a reference for how this can be done and should be used as a template as each division develops their own approach as is available at: https://earthdata.nasa.gov/collaborate/new-missions.

Status 1: Some open data requirements exist within SMD. The National Academies report provided recommendations as to how these requirements could be implemented more uniformly across divisions and phased in to maximize impact and minimize burden on the research community. Standard open data requirements should be developed over the next 12-18 months, in time to inform ROSES 2021. The Scientific Data and Information Officer should lead implementation.

Recommendation 2a: SMD should adopt a requirement for all new software development to be open source, except in instances of ITAR, EAR, national security, PII, or other similar restrictions. This is consistent with OMB-M-13-13. Any waiver to not having publicly accessible software will need to be approved by the Deputy Associate Administrator for Research.

Recommendation 2b: Any software developed under the ROSES NRA must be released as Open Source Software (OSS); opt-outs and alternatives must be justified and approved by the Deputy Associate Administrator for Research. The Earth Science Division currently has an OSS policy that could be adapted for a Directorate-wide policy[5]. This software policy must be added to the standard ROSES language. NPR 7120.8, and other NASA guiding documents, should be updated to include software in the compliance matrix.

Recommendation 2c: For new missions, an open source requirement must be included in the standard AO and SALMON language, and compliance should be evaluated as part of PLRAs and

[5] https://earthdata.nasa.gov/collaborate/open-data-services-and-software/esds-open-source-policy

appropriate lifecycle reviews. For missions already in operation, software accessibility should be part of the senior review process. NPR 7120.5, as well as the PE Handbook and relevant mission documents, should be updated to include software in the compliance matrix.

Status 2: This is not yet started and should be implemented by the Deputy Associate Administrator for Research in coordination with the Scientific Data and Information Officer.

AREA 2: High-End Computing (HEC) Program

Recommendation 3: SMD should conduct assessments no less than every five years for high-end computational resources. These assessments should evaluate HEC capacity, computing needs, and allocations across science divisions. This information should be used to develop a strategy for tracking and allocating cycle time for NASA resources and exploring whether there is the demand for additional or new computational resources to support SMD's objectives. The output of this assessment will also be used as part of Agency-level planning through the Office of Strategic Infrastructure's Shared Capability Assets Program to determine if additional resources are required.

Status 3: On August 7, 2019, the HEC Program made technical cloud computing services available to all NASA technical computing users. Training materials and sessions are available for users. A process will need to be developed within SMD to make this capability available to proposers in time for ROSES 2021.

The HEC Program plans to do an annual assessment that will rotate among mission directorates. The first assessment will be conducted in 2020 and will look at SMD. The HEC Program Manager will work with the Allocation Authority in each science division to prepare for the assessment.

The HEC Program is preparing a capability portfolio commitment agreement (CPCA) to be approved by SMD in Spring 2020, with the concurrence of the other mission directorates and OCIO. A draft has been developed and details are under negotiation, including additional research and development efforts in the areas of new system development/adaptation, algorithm development, code porting, and possibly bespoke system development, with a target signing date within the next six months.

AREA 3: Archives Modernization

Recommendation 4a: Digital data derived from NASA-funded research is required to go to a NASA archive for long-term curation and public availability. There are shared responsibilities between the principal investigators (PIs) and the archives to ensure transition of data. Archives must provide clear guidance on requirements to ingest data and investigators must adhere to the requirements. The Earth Science Division provides one example of such guidance at https://earthdata.nasa.gov/collaborate/new-missions. Additionally, the current archives must prepare to receive and validate data from heterogeneous sources, which may involve more "hands-on" work with PIs submitting data, metadata, and documentation. Such additional work will require additional funding, which should be addressed in the PPBE process.

Recommendation 4b: SMD should include an evaluation criterion for future ROSES solicitations to assess the adequacy of data management plans associated with individual proposals to support delivery of data to a final archive. SMD should evaluate whether an augmentation of funding may be required, either at the award or program element level, to enable this. As this becomes the norm, SMD should reassess the level of funding required.

Status 4: Archiving requirements currently exist for all missions at the Mission of Opportunity level or above, but not for activities solicited via ROSES or other managed under NPR 7120.8. Expansion of these requirements to make data discoverable and usable is under discussion within each division and an opportunity exists to coordinate across SMD.

The four divisions have independently started to evaluate data management plans, but this is not standardized across the entire organization. The Scientific Data and Information Officer shall work with the Deputy Associate Administrator for Research to develop standard criteria to inform ROSES 2021.

Recommendation 5: SMD should strongly encourage collaboration and cooperation of data professionals in academia, industry, and elsewhere to enable cross-cutting scientific discovery. There are currently small-scale efforts in the scientific communities to provide data stewardship training, including cleaning and curation, but these will need to be expanded to support anticipated future demand. This activity should be a sustained to ensure that new techniques (science or data), methods and algorithms for analysis, and management of data can be incorporated as they are developed.

Status 5: The community is currently ahead of SMD in this area. The Data Officer will need to engage with the relevant communities to understand how to leverage and enhance their efforts.

Recommendation 6: SMD should communicate the value of this capability and work with the NASA Centers to increase their recruitment of data science professionals. While NASA is at a competitive disadvantage for these types of positions, there are unique opportunities provided to data science professionals that only exist in the NASA environment.

Status 6: This is a conversation that SMD leadership will need to have with the NASA centers, potentially as part of the strategic workforce and ISFM discussion, as well as the NASA Office of the Chief Information Officer.

Recommendation 7: A process should be developed for data and software usability, discoverability, and accessibility to be evaluated as part of future senior reviews. This should also include utilization of advanced computing capabilities (e.g. high-end computing and commercial cloud computing). Information gathered through this process should be used to inform decision-making by the relevant archives.

Status 7: This recommendation will be implemented in time for the next Heliophysics Senior Review in 2020. In addition, it is under discussion for the upcoming Astrophysics Archives

Review in 2020. Lessons learned from these processes should be used to inform future similar reviews in other disciplines.

Recommendation 8: In order to enable cross-disciplinary science, SMD must facilitate greater discoverability of like data holdings in various archives. SMD should engage with the science community to develop a metadata description that would cross science boundaries in order to facilitate these new areas of study.

Status 8: This is a longer-term implementation activity that the Scientific Data and Information Officer should enable.

Recommendation 9: SMD should create a free and open, unified journal server along the lines of PubSpace, ADS or ERS to make science papers more accessible to the public. NASA Science should also consider adopting the National Science Foundation's requirements to reinstate the need for grant recipients to provide copies of all published research as part of their annual reports.

Status 9: ADS has provided estimates of the resources needed to include the relevant Heliophysics and Planetary Science journals. SMD leadership needs to make a decision on whether to accept this proposal or consider alternatives.

AREA 4: Advanced Capabilities

Recommendation 10: SMD should encourage the science divisions to explore novel computational techniques, including cloud computing and AI/ML, through various avenues, including ROSES NASA Research Announcements, technology calls in cooperation with the Office of the Chief Information Officer and Space Technology Mission Directorate, and mission Announcements of Opportunity.

Status 10: Earth Science and Heliophysics have undertaken a number of pilot projects to explore the use of cloud computing, AI/ML analysis techniques on very large data sets. Based on the interesting science results that this work has generated, efforts are being expanded. The Earth Science Division's data archival and access systems will begin full operations in the commercial cloud by January 2020. This environment can be used as a testbed for novel processing techniques applied to petabytes of data.

Recommendation 11: SMD should make investments to incentivize and educate the community on how to use AI/ML to approach science in new ways. Hands-on training can be achieved through expansion of hackathons, competitions, and grant programs. Science results and lessons learned about the use of AI/ML will be shared at community meetings to increase awareness of the potential of these techniques.

Status 11: Given the maturity of existing pilot programs, SMD leadership must determine how aggressively to pursue new opportunities utilizing these techniques.

AREA 5: Management

Recommendation 12: SMD should appoint a directorate-level Scientific Data and Information Officer to oversee this effort and serve as an interface between the Office of the Chief Information Officer, Office of the Chief Scientist, international partners, commercial providers, and others. The Scientific Data and Information Officer's responsibilities should include: 1. Setting policy for SMD, verifying compliance, and maintaining awareness of external policy drivers; 2. Maintaining a cross-divisional innovation forum to identify, select, and fund new opportunities; 3. Investing in targeted capabilities and tracking their progress; 4. Cultivating SMD's data and computing community through workshops, studies, training, etc., and; 5. Conducting periodic independent evaluation of the structure and content of the SMD data and computing portfolio.

Status 12: This is the central recommendation to being able to implement this strategy. SMD is in the process of hiring a Scientific Data and Information Officer.