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

Patricia Knezek, Program Scientist

Presentation to APAC

March 6, 2020

Background

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 was identified as a priority for assessment and action during the SMD Senior Leadership retreat in May 2017.

In February 2018, SMD chartered a working group with representatives from each division to develop a new SMD-wide strategy to enable greater scientific discovery over the next five years by leveraging advances in information technology to improve SMD science computing and data archives.

Today's presentation will cover the working group's findings and recommendations and planned next steps.

NAC SC Recommendations

Overall, NAC SC agrees with the BDTF that SMD data archive programs and projects are performing well and are properly taking steps to modernize. However, the volume, variety and velocity of NASA science data is taxing established methods and technologies. The SC finds that SMD should:

1) make investments in hardware, software, training and education to accelerate modeling workflows,
2) participate in the Department of Energy's (DOE) exascale computing program,
3) implement server-side analytics (SSA) capabilities (with caution),
4) forge a joint program with the National Science Foundation's (NSF) Big Data Innovation Regional Hubs and Spokes program, and,
5) incorporate data science and computing advisory positions in the SMD advisory committees.

In all efforts, the SC underscores that it is important that data science and computing experts work closely and collegially with domain scientists to implement effective solutions that are based on an understanding of the domain.

As to the future, the SC commends that an SMD Strategic Data Working Group has been set up that will bring forward these ideas, without interfering with how each division manages data.

For more info, reference "NAC SC Big Data Product" that contains SC feedback on each of the SC's Ad Hoc Task Force on Big Data findings/recommendations: https://science.nasa.gov/science-committee/meetings

NAS Open Code Recommendations

Recommendation: NASA Science Mission Directorate should develop internal policies and external legal language conducive to the swift release of open source scientific software, and the full participation of NASA employees in internal and external open-source projects, without jeopardizing national security or incurring legal liability.

Recommendation: NASA Science Mission Directorate should encourage the use of standard open source licenses, but not mandate a particular license. Non-standard licenses should be justified in the software management plan.

Recommendation: Any open source software policy that NASA Science Mission Directorate develops should not impose an undue burden on researchers; therefore, any policy should be as simple as possible and fully fund any mandates

Recommendation: NASA Science Mission Directorate should consider a variety of policy options depending on discipline and software type and transition to greater openness over time.

Recommendation: NASA Science Mission Directorate should support the infrastructure, governance, and maintenance of a healthy open source community, taking advantage of existing community resources to the greatest extent possible.

Recommendation: NASA Science Mission Directorate should support open source community-developed libraries that advance NASA science.

Recommendation: NASA Science Mission Directorate should foster career credit for scientific software development by encouraging publications, citations, and other recognition of software created as part of NASA funded research.

Recommendation: NASA Science Mission Directorate should initiate and sponsor programs to educate and train researchers in open source best practices. Topics could include, but are not limited to export controls, licensing and intellectual property, workflows, and software development. These resources could be made available to the community via in-person trainings as well as webpages, screencasts, and webinars.

Recommendation: NASA Science Mission Directorate should explicitly recognize the scientific value of open source software and incentivize its development and support, with the goal that open source science software becomes routine scientific practice**.**

Approach to Developing Strategy

As defined by the working group, the SMD 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 effectiveness 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

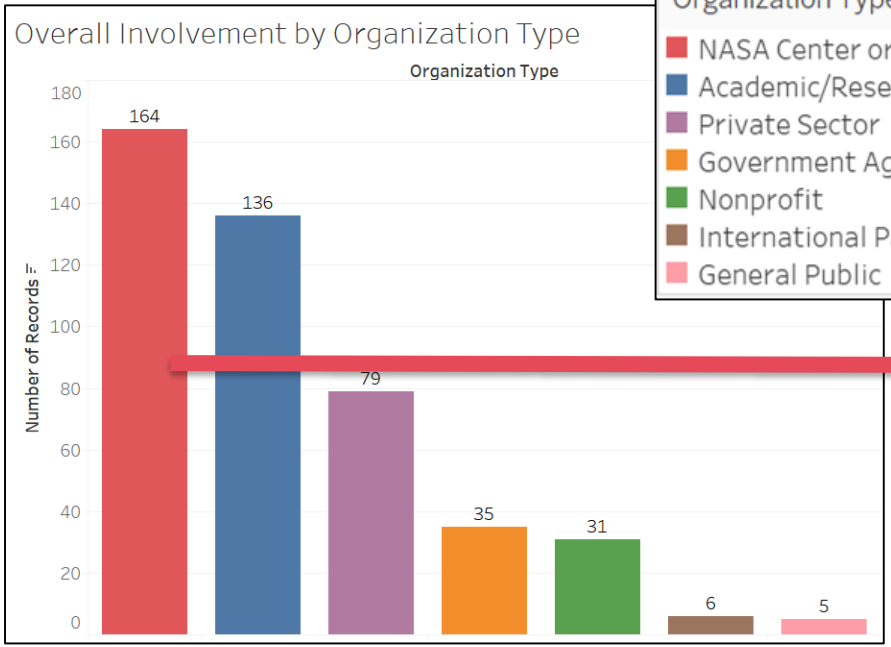
Approach to Developing Strategy (2)

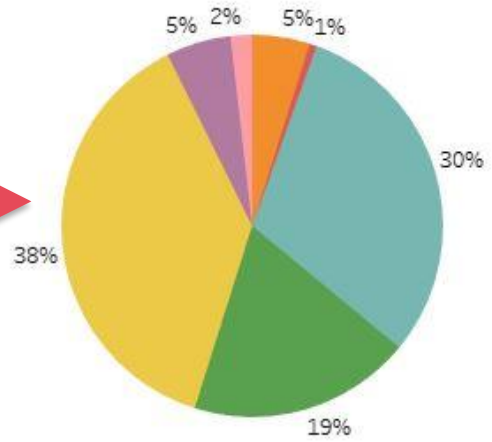
Given the breadth and potential impacts across the SMD community as a result of this strategy, the team used several mechanisms to collect stakeholder feedback and to promote data sharing and information gathering:

- Archives Processing and Data Exploitation Meeting, August 9-11, 2018
- NAC Science Committee draft analysis and findings in response to the report of the BDTF, August 28, 2018
- NASEM *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

Note: Information on all of these activities, and the report itself is at: https://science.nasa.gov/researchers/science-data

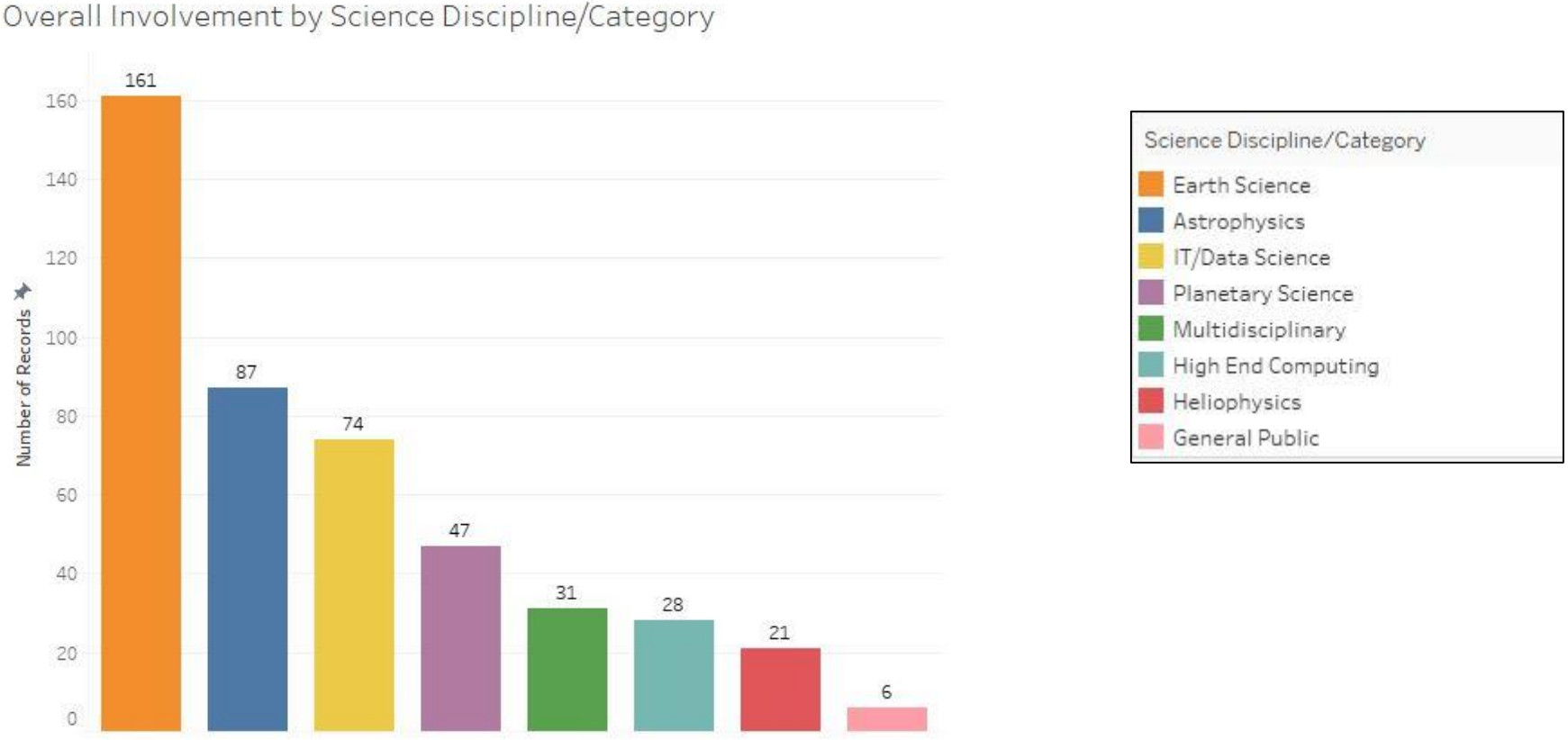
Stakeholder Participation by Organization





Over 450 people from academia, other Federal government agencies, foreign space agencies, commercial providers, professional societies, and the general public shared their ideas with NASA to inform the new SMD-wide Science Plan.

Stakeholder Participation by Discipline



Core Tenets/Guiding Principles

- The status quo will not work. The rate of change in this area exceeds our current capacity and our current systems are not set up to allow us to be aspirational in the next five years without significant investment
- Create a strong foundation that enables SMD to be responsive as the field changes
- Centralized constraint model that sets policies for all of SMD, deals with exogenous risks/opportunities, and shares best practices across the entire community – Managed by an SMD Data Officer
- Do not want to go to a fully centralized approach – Consistent with the recommendation of the NAC Science Committee, divisions must be responsible for the specialized components and implementation of policies to meet the needs of their communities
- May want to reassess this model over time – Periodic evaluation, considering the needs and best practices in the future, is appropriate. Divisions also should be encouraged to conduct similar reviews over time to evolve their specialized capabilities.
- Management of equipment and hardware can be centralized.
- Want people who know the data to manage the data where it sits

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.		
Goal 1: Develop and Implement Capabilities to Enable Open Science	**Goal 2**: Continuous Evolution of Data and Computing Systems	**Goal 3:** Harness the Community and Strategic Partnerships for Innovation
Strategy 1.1: Develop and implement a consistent open data and software policy tailored for SMD	**Strategy 2.1**: Establish standardized approaches for all new missions and sponsored research that encourage the adoption of advanced techniques	**Strategy 3.1:** Develop common metadata standards for all NASA science data
Strategy 1.2: Upgrade capabilities at existing archives to support machine readable data access using open formats and data services	**Strategy 2.2**: Integrate investment decisions in High-End Computing with the strategic needs of the research communities using this capability	**Strategy 3.2**: Utilize the full capacity of advances in High-End Computing to achieve SMD's research goal
Strategy 1.3: Develop and implement a SMD data catalog to support discovery and access to complex scientific data across divisions	**Strategy 2.3**: Invest in capabilities to use commercial cloud environments for open science	**Strategy 3.3:** Promote opportunities for continuous learning as the field evolves through collaboration
Strategy 1.4: Increase transparency into how science data are being used through a free and open unified journal server	**Strategy 2.4**: Invest in the tools and training necessary to enable breakthrough science through application of AI/ML	
	Strategy 2.5: Cultivate a strong community of practice across SMD, the science archives, and the broader research community	

Findings and Recommendations

- Open Data/Open Software

  1a. Standard open data policy for all new missions
  1b. Standard open data requirements in ROSES
  1c. Lifecycle approach for long-term data curation after KDP-F
  2a. Open source software requirement for new software development
  2b. Open source software requirement for ROSES
  2c. Open source software requirement for new missions

- High-End Computing (HEC) Program

  3. HEC assessments no less than every five years

Findings and Recommendations (2)

- Archives Modernization

  4a. All digital data required to go to a NASA archive for long-term curation and public availability
  4b. Evaluation criteria for future ROSES solicitations to assess the adequacy of data management plans
  5. Collaboration and cooperation of data professionals to enable cross-cutting scientific discovery
  6. Work with the NASA Centers to increase their recruitment of data science professionals
  7. Data and software usability, discoverability, and accessibility to be evaluated as part of future seniors reviews
  8. Facilitate greater discoverability of like data holdings in various archives
  9. Create a free and open, unified journal server to make science papers more accessible to the public

Findings and Recommendations (3)

- Advanced Capabilities

  10. Explore AI/ML and other novel computational techniques through various avenues
  11. make investments to incentivize and educate the community on how to use AI/ML

- Management

  12. Appointment of an SMD-level Data Officer
      - If you know anyone who might be interested, we hope to have the job listing posted to USAJobs within the next month.

Changes to ROSES-2020

Additions to ROSES-2020 in Support of Open Data/Open Models/Open Code

- Software created as part of a ROSES award, whether a stand-alone program, an enhancement to an existing code, or a module that interfaces with existing code, will be made publicly available
  - within 90 days of the end of an award
  - when it is practical and feasible to do so, and when there is scientific utility in doing so.
- Software whose utility is significantly outweighed by the costs to share it, is not expected to be made available

Changes to ROSES-2020 (cont.)

Increasing Access to the Results of Federally Funded Research

- When a DMP is required, the sufficiency of the data management plan will be part of the grade given to the proposal and will have a bearing on whether or not the proposal is selected
- Even if a DMP is not required with the proposal, if peer-reviewed publications result from the award then any data behind figures or tables must be available electronically at the time of release

Changes to ROSES-2020 (cont.)

New Placeholder Program Element in Appendix E

- E.7 SUPPORT FOR OPEN-SOURCE TOOLS, FRAMEWORKS, AND LIBRARIES
- NOTICE: NASA intends to solicit research proposals for this element under ROSES-2020. The final text will be released in March as an amendment with a proposal submission deadline no earlier than July and no fewer than 90 days after the release of the amendment

ROSES-2020

Reminder (not new):

- Awards deriving from ROSES include terms and conditions requiring that as accepted manuscript versions of peer-reviewed publications (hereinafter "manuscripts") that result from ROSES awards be uploaded into NASA’s part of the PubMed Central (PMC) repository called NASA PubSpace
- Not doing so may delay or prevent awarding of funds

Placeholder Slide - SMD-level Data Officer

BACK UP

Strategic Data Management Working Group

Name	Affiliation
Ellen Gertsen, Co-Chair	SMD Front Office
Kevin Murphy, Co-Chair	ESD
Hashima Hasan	APD
Jeffrey Hayes	HPD
Pat Knezek	APD
Bill Knopf	PSD
Janet Kozyra	HPD
Tsengdar Lee	ESD
Jared Leisner	HPD
Rebecca McCauley Rench	PSD
Viet Nguyen	JASD
Mariel Borowitz*	SIMD

*IPA ended in Fall 2018

Software included in ROSES-2020

Short Name	Name	Description	Examples
Libraries	Libraries and toolkits	Generic tools implementing well-known algorithms, providing statistical analysis or visualization, and so on, which are incorporated in other software categories.	Numerical Recipes, NumPy, general FFTs, LAPACK, scikit-learn, AstroPy, GDAL
Analysis software	Analysis, post-processing, or visualization software	Generalized software (not low-level libraries) used to manipulate measurements or model results to visualize or gain understanding.	Stand-alone image processing, topology analysis, vector-field analysis, satellite analysis tools, and so on
Frameworks	Modeling frameworks	Multicomponent software systems that incorporate a variety of models and couple them together in a complex way.	Community Earth System Model (CESM) is a collection of coupled models including atmospheric, oceanographic, sea ice, land surface, and other models