

National Aeronautics and  
Space Administration



# EXPLORE SCIENCE

Heliophysics Data Libraries:  
- *a new paradigm for research*

Jeffrey Hayes (jeffrey.hayes-1@nasa.gov)

Heather Futrell (heather.a.Futrell@nasa.gov)

Patrick Koehn (patrick.koehn@nasa.gov)

Dominique Chamely (dominique.m.chamely@nasa.gov)

*March 31, 2020*



A vibrant space-themed background featuring a large blue planet with a ring system, a smaller brown planet, and a bright yellow sun. The scene is set against a backdrop of a blue and green nebula with numerous stars. A large, light blue curved shape frames the right side of the slide.

# Outline

- Current State of Archives and CCMC
- Driving Policies, White Papers, and Responses
- Strategic Working Group
- Next Steps



# Current State: Where We Are and Why Now?

All disciplines have seen an explosion of data holdings over the last decade:

- Driven by inexpensive compute and storage, and the digital collection of data.
- PetaByte datasets are becoming the norm, not the exception: some disciplines are seeing >100 ExoByte datasets.
- Simple *archiving* is no longer practical: curation and the acquisition of metadata\* are essential to maintain the long-term investment that has been made to obtain data (Think > 100 years).
  - \*Metadata (def): how the data was calibrated analyzed; the description of the algorithms that did that calibration and analysis; and the science conclusions made from said analysis.
- For government funded research, all of this should be (and is now required to be) *publicly* available.



# Current State: Where We Are

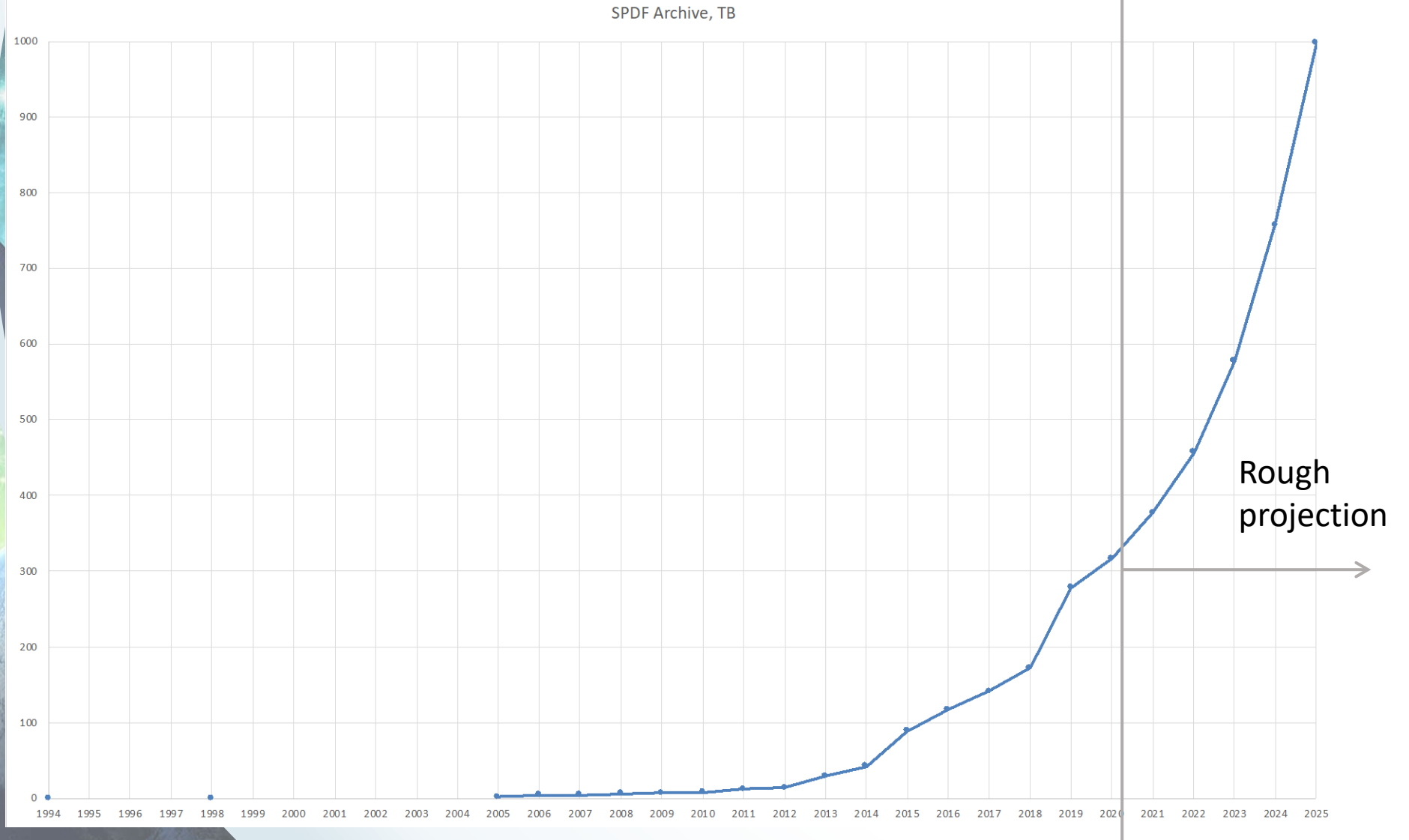
The NASA Heliophysics Division currently has two archives, one competed line for data recovery/data upgrades/code production, and a modeling center allowing the community to run state-of-the-art models on demand.

The division also has a written *Data Policy* that *requires* all data to be generated and submitted to archives in *standard* formats. There is also a metadata standard that all data are described in. The use of metadata allows for the creation and maintenance of a *registry* to make all data holdings discoverable.

NASA Heliophysics is moving to open code/open algorithms by supporting Python and other activities that will create libraries of standard analysis routines that can be used by the community as a whole (see ROSES 2019/HDEE call).

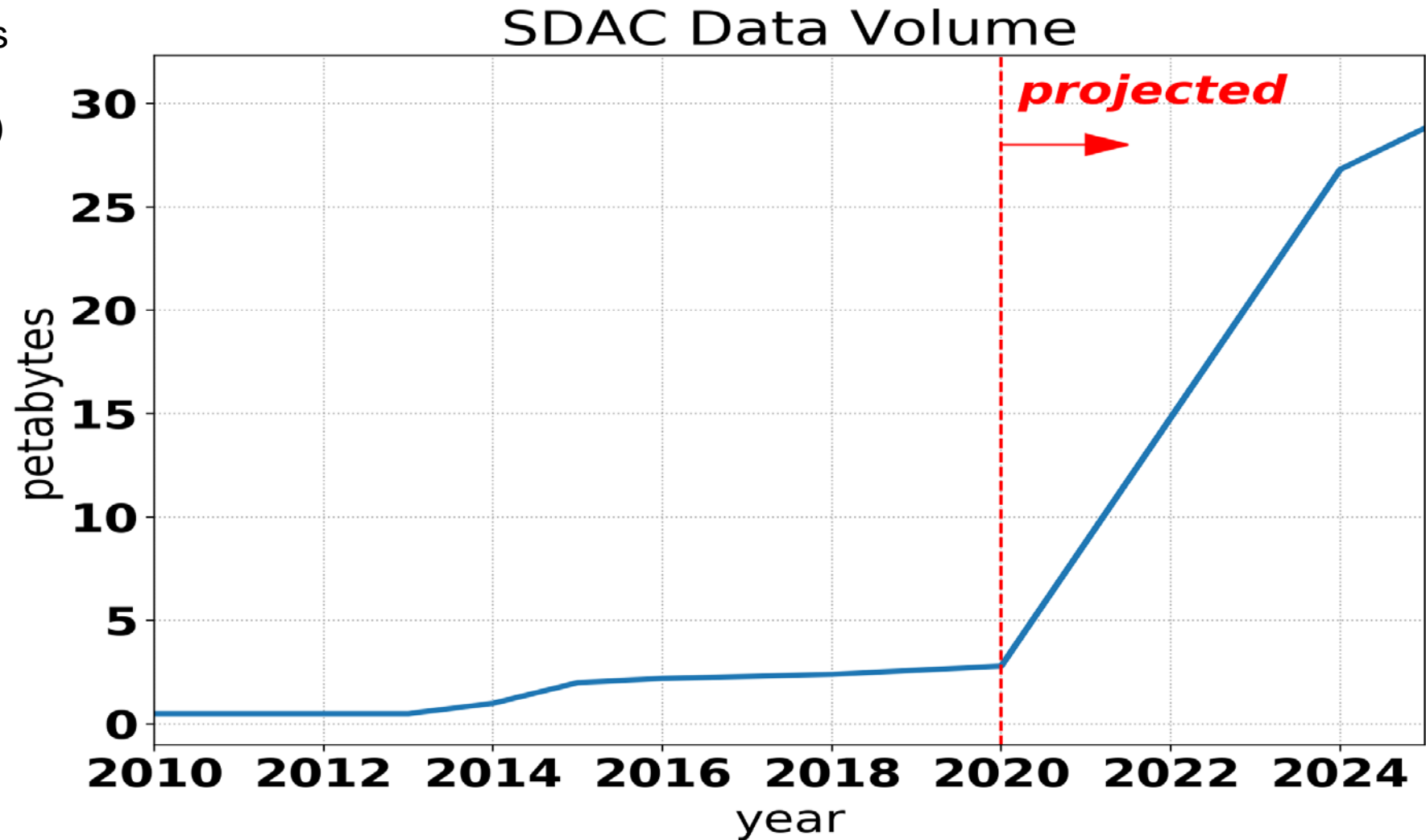
# Current State: Why now?

NASA's Space Physics  
Data Facility's data holdings:  
2010 – 2025 (TB).



# Current State: Why Now?

Solar Data  
Analysis Centre's  
Data holdings  
2010 - 2025 (PB)







# Current State: Why Now?

From the previous plots, it is clear we are in an era of unprecedented data growth.

There are some practical implications on this exponential trend:

- Conventional storage and retrieval becomes impractical;
- Simple archiving, with the aim of discoverability and usability cannot be maintained in the *status quo*;
- *Curation*; (i.e. the long-term preservations of data, metadata, and other products) *needs to become the new norm*. Saving the bits isn't enough in this era. This implies new skillsets for archivists;
- To maximize the usefulness of the data, and show the taxpayer investment was worth it, data need to be made public.

HPD started a pilot project of placing the entire OMNIWeb database from SPDF into the Amazon Cloud. Initial lessons learned are that: 1) data needs to be made 'Cloud' ready; 2) even with imperfectly sampled data, AI/ML algorithms are finding super-storm events that have never been seen before – an unexpected and very exciting result that shows the power of these techniques.

# Current State:

## Solar Data Analysis Center (SDAC)

- Current data volume: 2.6 PB (~15 PBs coming)
- Virtual Solar Observatory (VSO), helps distribute data from SDO AIA and HMI (another ~4PB).
- In addition to management of archives, SDAC provides management of:
  1. Virtual Solar Observatory development and maintenance
    - <https://sdac.virtualsolar.org/cgi/search>
  2. SolarSoft library development
    - [http://www.lmsal.com/solarsoft/sswdoc/sswdoc\\_jtop.html](http://www.lmsal.com/solarsoft/sswdoc/sswdoc_jtop.html)
  3. Helioviewer development and operations
    - <https://helioviewer.org/>



**Solar Data Analysis Center**

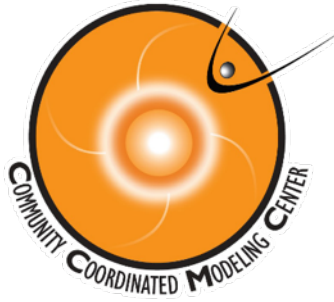


# Current State: Solar Physics Data Facility (SPDF)

- Current data volume: ~100 TeraBytes
- Other measures of SPDF scope
  - ~90 million total archived files
  - ~100 spacecraft/sources in CDAWeb, >1400 (CDF) datasets
  - >200 spacecraft in SSCWeb and available in 4D Orbit Viewer
  - ~2400 datasets now described in HDP
- Latest missions that will produce their data in netCDF:
  - Solar Orbiter, Parker, ICON, and GOLD

*Space Physics Data Facility*





# **Current State: Community Coordinated Modeling Center (CCMC)**

## **Project Description:**

- Facilitate development, host and execute next generation space environment models in support of the advancement of space science and development of operational space weather prediction capabilities.

## **Science Goals:**

- To model and understand the physical processes of the space environment throughout heliosphere.

## **Partnerships:**

- CCMC is a partnership with NSF, NOAA, AFRL, AFOSR, AFWA, AFMC, and ONR
- Models hosted at CCMC are developed by the international research community, and the tools and services are available world-wide: all are heavily used and acknowledged all over the world.

# **Evolving Policy is Moving to Open Data**

**The era of Big Data is upon us, and we as a discipline need to both take advantage of it and embrace it. (NASA is obliged to do so...)**

## **OMB:**

- M-13-13: Open Data Policy-Managing Information as an Asset (2013)
- M-16-16: Agency Open Government Plans (2016)
- Public Law 115-435: Foundations for Evidence-Based Policymaking Act of 2018 (2019)

## **National Academies of Science, Engineering and Medicine:**

- Open Science by Design: Realizing a Vision for 21st Century Research (2018)
- Open Source Software Policy Options for NASA Earth and Space Sciences (2018)
- Advancing Open Science Practices: Stakeholder Perspectives on Incentives and Disincentives: Proceedings of a Workshop in Brief (2020)



# NASA's Response to Data Management Policy Challenges

## **NASA Agency:**

- NASA Plan for Increasing Access to the Results of Scientific Research (2014).

## **Science Mission Directorate:**

- Chartered Strategic Data Management Working Group (2017) – more below.

## **Heliophysics:**

- NASA Heliophysics Data Policy: Current Version 1.2 of 2016 is being updated.
  - B.1 The Heliophysics Research Program Overview (used for ROSES Research and Technology, ROSES 2020) updated to include Project Data Management Plans (PDMPs);
  - PDMPs now required in all Senior Reviews (introduced in Senior Review, 2020).

# Data Management Working Groups

## SMD Strategic Data Management Working Group

- Started in 2017 by SMD Leadership:
- Membership across all science divisions.
- Objective: Enable greater scientific discovery by leveraging advances in information technology for SMD science data archives.
  - Recommendations made to and accepted by SMD management Dec 2019.
  - **SMD currently drafting high-level policy for implementing data accessibility and usability.**
    - *Open Source Algorithms and Code Policy to come later this year*

## Heliophysics Archives Strategic Working Group

- Started in 2019
- Established to assess, restructure, and modernize the HPD Archives
  - Meant to complement and augment the work being done at the SMD level.
  - Per SMD draft data policy, *implementation* is left to the individual divisions.


# HPD Archive Strategic Working Group

In line with US government strategy, the NASA Heliophysics Division (HPD) Archives are committed to being the premier resource for all NASA HPD data needs. Moving beyond a traditional repository and toward a functional, collaborative data library, the NASA HPD archives will maximize the utility of the data of the HSO, sustainability of the archives, and access for the public to this data.

## **Vision:**

*Democratize the  
Science & Data of  
Heliophysics*

## **Mission Goals:**

- 
- Provide the infrastructure for a functional HPD data library
  - Identify and support data providers
  - Curate Integrated Heliophysics Data
  - Serve the Public as a working Data Library
  - Maximize the Utility of the data
  - Expand the field's engagement with NASA HPD data



# Working to a *Heliophysics Data Library*

**Infrastructure:** Exploring new ways to store and access data. A Cloud-like architecture is likely, but would it be private or public/commercial? What does egress look like? Computation in the Cloud?

**Data Providers:** Need to ensure that providers know expectations for deliverables and have access to the tools and expertise to make usable data products.

**Curation and Service:** NASA needs to infuse the techniques and best practices of long-term curation (>100 years) into the system. This implies additional skill-sets (people) being need wherever the data resides. This is a very different way of approaching to what we have been doing till now.

**Maximizing Utility:** The Data Library ideally will be a resource center for the broad community with SMEs who can aid non-expert researchers in their studies.

## **Engagement:**

- Community engagement: Updates and news on data holdings and tool development. Creation of a Users' Working Group.
- Interagency cooperation: Coordination with NFS, NOAA, ESA, JAXA, ... on accessibility of all data to create an international data library.
- The NASA holdings are ideal for mining via *Citizen Science* activities and other public, hands-on research experiences.

# HPD Data Library Working Group

## Next Steps:

- Continue working to modernize the archives into a Heliophysics Data Library (HDL)
- Focus topics include:
  - Data Curation (greater than 100 years)
  - Open Source Code and algorithms
  - Accessibility and interoperability between data and models
  - Machine Learning/Artificial Intelligence to enable new science
  - High End Computing needs for modeling
- Establishing an HPD Data Library Working Group (HDLWG) with more stakeholders to ensure the true big picture is understood and the new HDL will, and continue to, meet the needs of its users

**Volunteers Welcome**