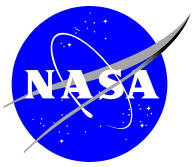
National Aeronautics and
Space Administration



Heliophysics Archiving Strategy.

Vision:

Democratize the Data & Science of Heliophysics

Mission:

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 Heliophysics System Observatory (HSO), sustainability of the archives, and access for the public to these data.

Background:

All science disciplines have seen an explosion of data holdings over the last decade. This has been driven by inexpensive computing and digital storage, and the digital collection of data.

PetaByte datasets are becoming the new norm, not the exception; indeed, some disciplines are seeing in excess >100 ExoByte datasets. The simple *archiving* of data is no longer practical – the curation and the acquisition of metadata are essential to ensure the longevity (>100 years) of data usability and usefulness given the investment made in each mission.

For government funded research, all of these data should be, and are now required to be both fee-free and *publicly* available.

Goals:

Evolving and adapting to support the current and future community needs, HPD seeks to create a system where all NASA-funded (as well as complementary domestic and international datasets as appropriate) data will be available via a searchable interface and made available and accessible to the broadest audience, including government, international partners, the research community, and the general public. The requirements for open source, open access, and open development with regard to data will be observed.

Archives Strategy by Goal				
	Theme	Goal	Objective	
	1. Develop	Provide the infrastructure for a functional	1.1	Establish standards, protocols, procedures, documentation, equipment/software, and technology as needed
			1.2	Maintain said standards, protocols, documentation, equipment/software,

		Heliophysics Digital Resource Library (HDRL)		technology, simulation services, and analysis services as needed
			1.3	Upgrade hardware and software to match the technology as it evolves
			1.4	Ensure adequate staffing, training, and regular review of required capabilities for HPD Data Library Curators
			1.5	Heliophysics Data Archive working group
			1.6	Ensure comprehensive datasets including CubeSats, SmallSats, sounding rockets, and balloons
			1.7	Coordinate with and across SMD regarding archival efforts
	2. Unify	Identify, connect, and unify support for data providers	2.1	Identify, build, maintain relationships with new, existing, and former providers
			2.2	Provide technical assistance to missions and the data provider community to build capacity and ensure adherence to standards
			2.3	Engage and support synthetic data providers, model output providers, and model developers
	3. Curate	Curate integrated heliophysics data	3.1	Ensure the long-term preservation of heliophysics datasets (i.e., 100+ years)
			3.2	Acquire/receive the data via a standardized process
			3.3	Adhere to FAIR principles and processes for existing and new datasets.
			3.4	Ingest the data according to the terms of the PDMP
			3.5	Establish and regularly update quality assurance protocols for curated data
			3.6	Maintain a standard process for a deep "backup"/cold storage
	4. Operate	Serve the Public as a working *Digital Resource Library*	4.1	Staff research librarians to scope and enhance data products
			4.2	Design and host a public-facing website/interface for viewing all the data
			4.3	Design guides to understand how to access and use the data
			4.4	Provide support to stakeholders
			4.5	Produce and maintain data access via an application programming interface
	5. Optimize	Maximize the Utility of the data	5.1	Continue to host community-provided models and provide access to modeling services
			5.2	Conduct in-house analysis
			5.3	Partner with HPD Citizen Science
			5.4	Baseline and continually track metrics
			5.5	Utilize data-informed analytics to identify opportunities enhancing services for Archives Stakeholders
			5.6	Establish and maintain access to and analysis services for heliophysics data products
	6. Grow	Expand the field's engagement with NASA HPD data	6.1	Communicate the goals and objectives of the HDRL for existing and new stakeholders
			6.2	Foster peer-to-peer connectivity between members of the data provider community

1. **Goal 1: Provide the infrastructure for a functional Heliophysics Digital Resource Library (HDRL)**

Justification: The Heliophysics community is in a time of unprecedented data growth. There are practical implications on this exponential trend:

- conventional storage and retrieval become impractical due to the sheer sizes of the datasets involved;
- simple archiving of data and software is inadequate in this era; in order to maximize the return on investment of the data, the findability, accessibility, interoperability, and re-usability of the data must be maximized
- data and software must be open and accessible to the public; and

- *curation* (i.e., the long-term preservation of data, metadata, software, and other products) *needs to become the new norm.*

Due to all these reasons, there is a need for new skillsets for archivists and users of these data holdings.

Objectives:

1.1. Establish standards, protocols, procedures, documentation, equipment/software, and technology as needed

1.2. Maintain said standards, protocols, documentation, equipment/software, technology, simulation services, and analysis services as needed:
- Standardization of data level product formats produced and archived
- Maintain and set protocols for the Space Physics Archive Search and Extract (SPASE) metadata standard
- Establish a "Registry Czar" role to oversee the minting of Digital Object Identifiers (DOI)

1.3. Upgrade hardware and software to match the technology as it evolves

1.4. Ensure adequate staffing, training, and regular review of required capabilities for HPD Data Library Curators
- Determine what data curation programs/certifications exist and if new ones need to be developed

1.5. Heliophysics Data Archive working group
- Quarterly review of resource utilization and prioritization of activities of import to the community and provide findings to NAS/HPD about the priorities

1.6. Ensure comprehensive datasets including CubeSats, SmallSats, sounding rockets, and balloons
- Hold discussions with other agencies concerning the archiving of ground- and laboratory-based research-based datasets.

1.7. Coordinate with and across SMD regarding archival efforts

2. **Goal 2: Identify, connect, and unify support for data providers**

Justification: With the increasing diversity and knowledge levels of data users and providers (missions, researchers of ROSES awards, etc.), it is essential that the HDRL be able to support and assist this community. This requires up-to-date knowledge of all data being curated in the Library, as well as knowledgeable subject matter experts (SMEs) to assist non-SME researchers and providers with timely and accurate help—this will not only build capacity within the data provided but also serve to ensure continuity of best practices.

Objectives:

2.1. Identify, build, maintain relationships with new, existing, and former providers, including but not limited to:
- CCMC, SPDF, SDAC -- intercommunication essential
- NASA Missions (past, present, future)
- PIs and other researchers
- international partners
- Other domestic agencies

2.2. Provide technical assistance to missions and the data provider community to build capacity and ensure adherence to standards by:
- Enacting Mission Commitment Agreements protocol (which is the initial agreement plus a commitment to participate in technical assistance and communicate with archives throughout the process);

- Negotiating Project Data Management Plans (PDMPs) by Key Decision Point (KDP)-C;
- Support with data design and quality assurance, including:
  - Formatting into acceptable international standards
  - Packaging with appropriate documentation and metadata
  - Overseeing DOI minting/registration
  - Delivery for archiving and curation
  - Determining data interface with ML/AI
- Monitor compliance both at the Library level, and reviewed by the Working Group, and programmatically by NASA/HPD
- Undertake Biennial Management Reviews (i.e., ensure that industry and community best practices are followed)

2.3. Engage and support synthetic data providers, model output providers, and model developers

3. **Goal 3: Curate Integrated Heliophysics Data**

Justification: While the focus of this activity is to ensure that NASA/HPD data and metadata are being curated with the most up-to-date methodology and practices, NASA's data holdings are not the only ones to consider. There are numerous national and international partners that have extensive data holdings essential to understanding the physical processes involved with the Sun and its extended atmosphere as it interacts with the Earth, other planetary bodies, and the interstellar medium. The HDRL needs to develop data format and metadata standards that will allow ground-based and other space-based data from non-NASA missions to be able to 'hook in' to the HDRL and allow those data to be discoverable and useable to all researchers in a transparent way.

Objectives:

3.1. Ensure the long-term preservation of NASA's Heliophysics datasets (i.e., 100+ years)

3.2. Acquire/receive the data via a standardized process
- Legacy Data: Generate scripting to simplify data uploads
- Recent Missions: Final data delivery
- New Missions: Interim and final data deliveries

3.3. Adhere to FAIR principles and processes for existing and new datasets
- Ensure that data are (F)indable, (A)ccessible, (I)nteroperable, and (R)eusable

3.4. Ingest the data according to the terms of the PDMP
- Account for data version control

3.5. Establish and regularly update quality assurance protocols for curated data
- Develop and maintain standard checklists for assessing and ensuring data quality based on in situ/remote sensing data

3.6. Maintain a standard process for a deep "backup"/cold storage

4. **Goal 4: Serve the Public as a working *Digital Resource Library***

Justification: The expectation for the next generation of researchers and users of NASA/HPD data are that these data *not* require highly specialized knowledge of where the data are and how they are described. As the communities that the HDRL supports become more diverse, and as cross-disciplinary research becomes the usual mode, it is essential that the HDRL evolves to meet those needs including a central website for ease of user access and data searches. Registries that allow for the discoverability of data hosted by other organizations and tools to facilitate the use of these disparate data holdings must be developed to maintain the exceptionally high productivity in the research community and allow the non-

expert user an ability to contribute to the field. In addition, the HDRL will need to provide expertise and tools to allow users to contribute the results of their research *into* the HDRL.

Objectives:

4.1. Staff research librarians to scope and enhance data products
4.2. Design and host a public-facing website/interface for viewing all the data
4.3. Design guides to understand how to access and use the data
4.4. Provide support to stakeholders
4.5. Produce and maintain data access via an application programming interface

5. **Goal 5: Maximize the Utility of the data**

Justification: The research community is moving away from a simple analysis of a simple set of data to produce a specific scientific result, to one that requires the use of many datasets and simulations, as well as sophisticated models, to gain insights into the complex and inter-connected physical processes that heliophysics studies. To allow this to flourish, the HDRL needs to develop standards and tools that allow for the comparison and incorporation of instrument measurements with assimilative models. This has the potential to vastly increase our knowledge of the physical processes involved, but also implies very large datasets coupled with very computer-intensive models and simulations.

Objectives:

5.1. Continue to host community-provided models and provide access to modeling services:
- Analyze data with standard tools, with assistance (as needed) from domain experts

5.2. Conduct in-house analysis:
- In-house domain expert/research librarians work with the community in the design and generation of high-level datasets and analysis tools

5.3. Partner with HPD Citizen Science
5.4. Baseline and continually track metrics
5.5. Utilize data-informed analytics to identify opportunities enhancing services for Archives Stakeholders
- Information from unique IP addresses will be combined with additional anonymized user-provided information that will allow the HDRL to better serve the needs of the professional and citizen science communities

5.6. Establish and maintain access to and analysis services for heliophysics data products

6. **Goal 6: Expand the field's engagement with NASA HPD data**

Justification: NASA/HPD has a wealth of data and as those holdings evolve from archives into the HDRL, this means that anyone interested in the science will be able to access these data and use them for studies at numerous levels of traditional and non-traditional study. The use of HDRL data will also allow the users to be confident that the data they are using are appropriately vetted and that analysis tools and algorithms are the 'gold standard' in the field. Aside from improving the efficiency for researchers, it allows all users to participate in the scientific process without the steep learning curve that we experience now. It should also lead to extensive cross-discipline research areas that will change the way we think of doing teaching, research, and outreach.

Objectives:

6.1. Communicate the goals and objectives of the HDRL for existing and new stakeholders

- Outreach
- Education
- Engagement
- Host engagement and capacity building events (such as annual code-only workshop, etc.)
- Charter a Working Group to provide advice to the HDRL on community and industry best practices: members to be nominated by NASA/HDP to serve on a rotation-limited (no more than 3 year) basis

6.2. Foster peer-to-peer connectivity between members of the data provider community

- Hosting activities to enable collaboration between the groups, including:
    - Seminars or workshops sponsored by HQ and run by the data centers that bring together the data community (users and providers)

Acronyms used in this document:

CCMC: Community Coordinated Modeling Center
DOI: Digital Object Identifier
FAIR: (F)indable, (A)ccessible, (I)nteroperable, and (R)eusable
HDRL: Heliophysics Digital Resource Library
HPD: Heliophysics Division
HSO: Heliophysics System Observatory
IP: Internet Protocol
KDP: Key Decision Point
NASA: National Aeronautics and Space Administration
PDMP: Project Data Management Plan
ROSES: Research Opportunities in Space and Earth Science
SDAC: Solar Data Analysis Center
SME: Subject Matter Expert
SPDF: Space Physics Data Facility