Heliophysics Advisory Committee HDRL Briefing (GSFC)

Brian Thomas, Robert Candey, Jack Ireland, Lan Jian,
Rebecca Ringuette, Tressa Helvey-Kalsulke

Oct 22, 2024

HDRL
HELIOPHYSICS DIGITAL RESOURCE LIBRARY

Core Features and Functionalities of HDRL

(HDRL: Heliophysics Digital Resource Library)

HDRL Org Chart

HP Data and Model Consortium / HDMC

Brian Thomas (PS),
Tressa Helvey-Kasulke (PM)

Overall management of the HDRL.
Registries, Metadata and DOIs for all digital resources; SPASE Data Model.
Heliophysics Data Portal (HDP)
Python and other software integration (PyHC).
Analysis and visualization services ((Py)SPEDAS, Autoplot).
Data upgrades and related services.
HelioCloud initiative with data and shared software environments.

All activities within the various HDRL components are interrelated.

Space Physics Data Facility / SPDF

Robert Candey (PS), Lan Jian (DPS)

Non-solar Data Final Active Archive for NASA (and other) missions.
CDAWeb data browsing and access; Web Service access.
OMNIWeb data production and serving.
SSCWeb and 4-D spacecraft orbit facility. Common Data Format.

Solar Data Analysis Center / SDAC

Jack Ireland (PS)

Solar Data Final Active Archive for Solar Dynamics Observatory and other NASA missions.
Virtual Solar Observatory data access.
Helioviewer. SolarSoft. SunPy.
High Performance Computing for NASA HP.

Key Collaborators

Center for HelioAnalytics

Open Science, AI/ML Development, Mission-Enabling Tech, Community Resources, User Testing

Community Coordinated Modeling Center

Data-model comparisons; Registry of models and output; "Kamodo" enabled visualization.

HDRL Holdings

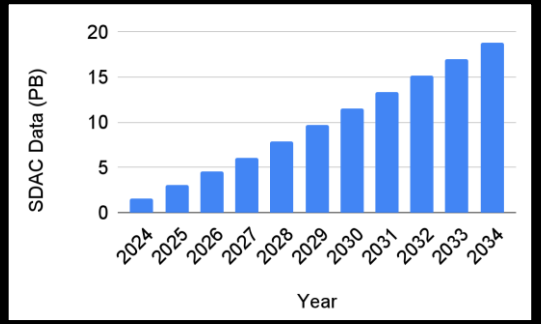
Some statistics (3Q FY24)

- ~300 million files
- 8000+ datasets
- 2.5+ PB of science data
- 30+ operating missions (& 100+ old missions)
- 600+ Instruments
- Over 65 years of data!

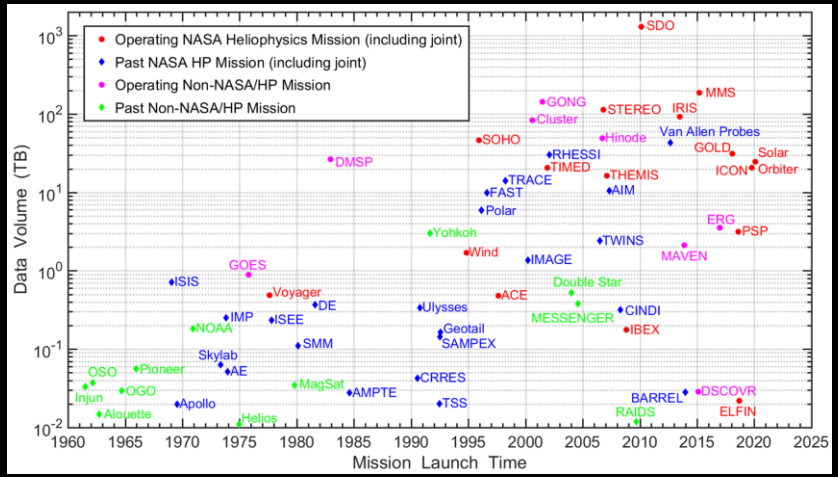
We have "Big Data"

Volume & Variety

System Science is not possible without HDRL



Projected Data Volume Growth to >20 PB



***HDRL data holdings** span 100+ missions over 65 years.*

How Does HDRL Ensure Accessibility and Usability?

HDRL ensuring accessibility and usability?

Curated by domain experts. Data oversight by member:

- **SDAC** (Solar Data Analysis Center) (Solar physics data)
- **SPDF** (Space Physics Data Facility) (Space Physics data)
- **HDMC** (data modeling/discovery/registration for solar- and space-physics combined)

Ensuring Usability

- ***Implementing Standard Formats & Metadata***
  - Solar : VSO metadata + FITS
  - Space Physics : ISTP metadata + CDF and netCDF
  - Both domains: SPASE metadata
- ***Engage with missions*** to deliver best science value.
  - Develop PDMP w/ each mission (e.g. set data format & metadata requirements, levels of service, etc)
- **Support community software** to accelerate use and analysis of the data
  - IDL: CDAWlib, SPEDAS, SolarSoft
  - Python: SunPy, PySPEDAS and PyHC (Python in Heliophysics Community, ~80 helio python projects), Autoplot
  - Data format support: CDF and support tools/libraries

Solar physics data may be accessed via the VSO.

__Logos__ of a some standards implemented by HDRL

SPEDAS and SunPy are examples of some of the software we support

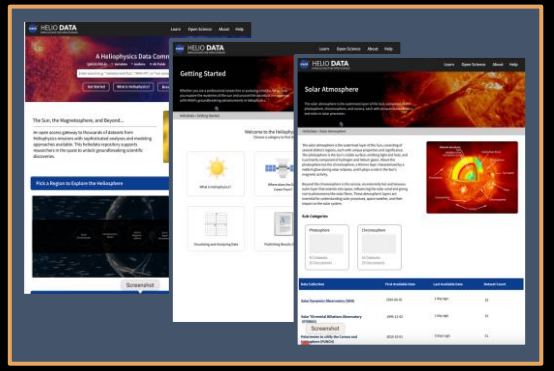
HDRL ensuring accessibility and usability?

We provide and support ***many*** discovery, access and visualization services for the community. Key services include:

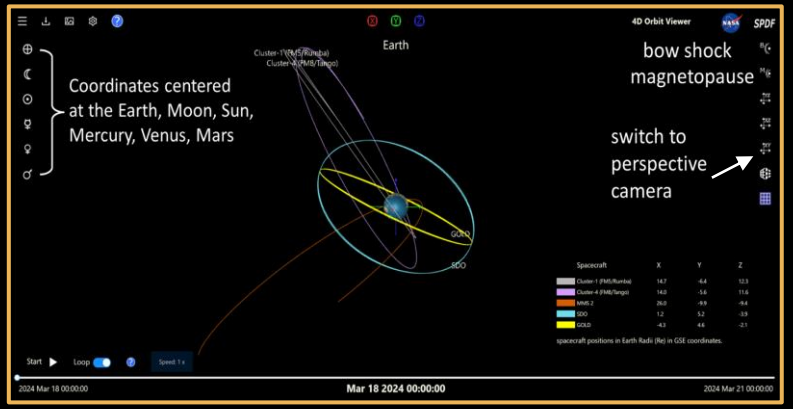
- Heliophysics Data Portal and API (HDP)
- CDAWeb, OMNIWeb and SSCWeb
- Virtual Solar Observatory (VSO)
- Helioviewer (2 versions: student and researcher)
- Heliophysics Event Knowledgebase (HEK)
- 4D Orbit Viewer
- helio.data.nasa.gov website (*new in FY25*)

Also:

- Implementing standard community & public protocols for access: HAPI, FTP, HTTPS, RESTful APIs
- 1 PB of science data on the cloud



Helio.data.nasa.gov. *Will feature improved interface and discovery capabilities (FY25)*



Improved 4D Orbit Viewer. *Features 160+ spacecraft.*

How Does HDRL Integrate With Other Data Repositories and Libraries?

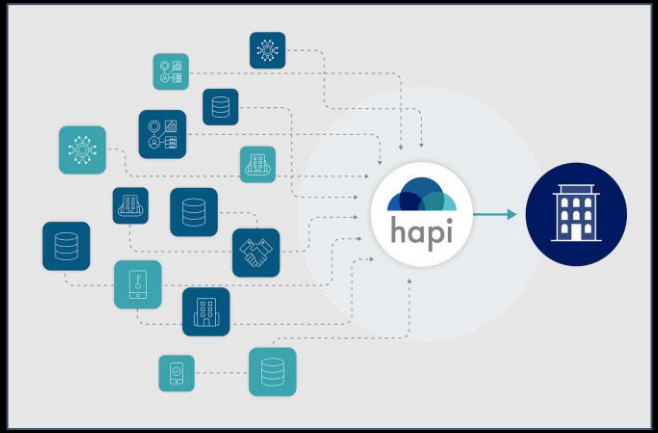
HDRL integration with other data repositories and libraries?

Engage partners in national & international community forums, meetings

- **IHDEA** : International Heliophysics Data Environments Alliance
- **IHDWG** : Interagency Heliophysics Data Working Group (US agencies)
- **VSO**: International coalition of solar repositories
- Many others including COSPAR, ESIP, etc.

Tangible and ongoing work with national and international partners to develop shared standards

- **Data formats & metadata:** FITS, CDF, SPASE, ISTP metadata, schema.org, DataCite (DOIs), Heliophysics software discovery.
- **Shared heliophysics software analysis environments** (w/ ESAC and others)
- **PyHC** community software standards and best practices (in Python)
- **HAPI** : Heliophysics "API" for standardized access across all heliophysics repositories

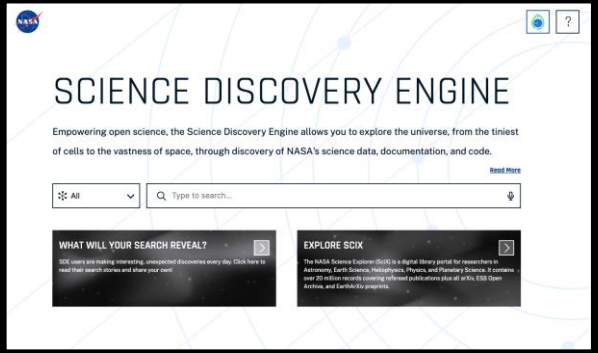


HAPI – Heliophysics API*. A COSPAR standard for accessing heliophysics resources. > 10,000 datasets are accessible from 10 servers. Clients are supported in 7 programming languages.*

HDRL integration with other data repositories and libraries?

Partner integration

- **SPDF and SDAC Services** : called by many international data and visualization services
- **OCSDO Science Discovery Engine** (via SPASE)
- **ADS/Sci-X integration** (interlinking and tagging datasets and papers with AI)



NASA's Science Discovery Engine. *HDRL provides over 2000 datasets.*

Sci-X dataset discovery*. Sci-X is now linking to HDRL datasets from papers which use these data.*

HDRL Outreach and Community Engagement

HDRL Outreach and Engagement

Meeting Outreach

- AMS, GEM, SHINE, CEDAR, TESS, AGU
- IHDEA, DASH

Workshops / Focus Groups

- Software Workshop (May 2024)
- Infrastructure Workshop (May 2023)
- PyHC Summer School (>500 registrations!)
- PyHC Infrastructure Workshop (Sept 2023)
- 20 questions exercise (Summer 2021)
- SMD Data Repositories Workshop

Upcoming

- Analysis Working Groups
- Summary report of outreach takeaways

Data, Analysis, & Software in Heliophysics. *HDRL is a founding member. An international EXPO of heliophysics data environment work.*

IHDEA website. *HDRL is a founding member of IHDEA which helps lead the international community in developing and maintaining standards for helio data environments.*

Selected Outreach Takeaways & Actions

Community is very happy with our services & support, BUT (since you asked…)

- **Open science needs infrastructure support**
  - It's expensive for researchers to do manually
  - Flexibility, automation, and streamlining needed for missions and researchers
- **Users want more advanced discovery**
  - Improved discovery interface
  - Search by phenomena, methodology, related software
  - Interlinking of research papers using data, related software
- **Users need big data support**
  - Democratizing High End Computing
  - Provide systems which allow for big data computing
  - Software environments which are 'user friendly'

→ **Open Science Hire**, developing strategy

→ **New Helio.Data** website (more interlinked resources, user friendly)
New Development of advanced discovery, semantics
Interlinking with ADS, use of OpenAI to rapidly link

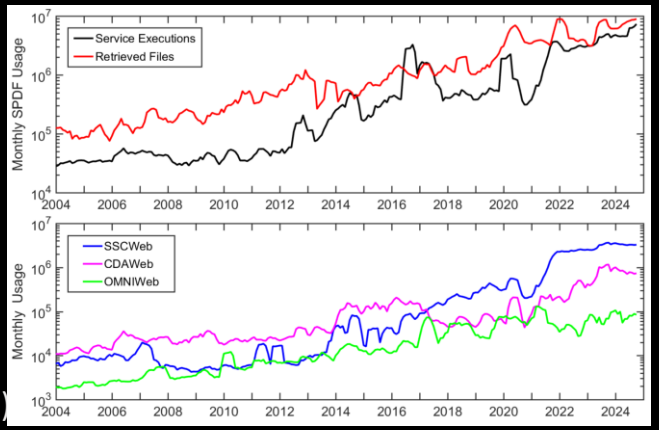
→ **HelioCloud** : Open architecture in cloud for heliophysics
On premise computing cluster at GSFC (non-NASA access possible)
Shared software environments (virtual env. & containers)

HDRL Impact

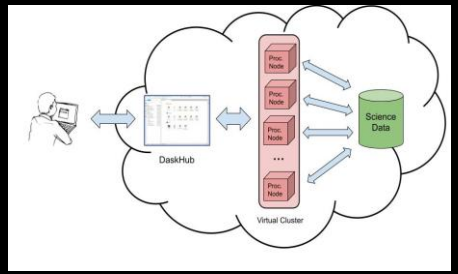
HDRL impact on research and education?

There are many areas of positive impact:

- **Highly utilized services** (SPDF, VSO)
  - VSO served an average of 1 TB/day of SDO data in August 2024.
  - Exponential usage growth (SPDF)
- **JGR and Space Weather publications** (50% papers reference use of SPDF data)
- **Supporting scientific community, citizen science, students & eclipse events**
  - Helioviewer users made over 380,000 movies in the last year
  - Eclipse service. Over 10,000 “eclipse” images served.
    - Launched only a week before the eclipse
    - Still available due to feedback from outreach team - easy to use, and simple idea.
- **Support for Frontier Development Lab** challenges and Machine Learning dataset hosting
- **HelioCloud support for research & various student workshops** (MUREP interns, Nepal student hackathon, PyHC summer school)



SPDF services have experienced exponential growth over past 2 decades.



Example of “Bursting” compute in HelioCloud. *~2-3 orders of magnitude faster processing. Over 180 researchers have accounts in HelioCloud and 10 research papers were written using HelioCloud in FY24.*

End of Presentation

Thank you!

HDRL Members

Federation of 3 member organizations

- **SDAC** : Solar Data Analysis Center (Jack Ireland)
- **SPDF** : Space Physics Data Facility (Robert Candey, Lan Jian)
- **HDMC**: Heliophysics Data & Modeling Consortium (Brian Thomas, Tressa Helvey-Kasulke)

Roles:

- SDAC : solar physics support
- SPDF : space physics support
- HDMC : provides integration and unified strategy between solar and space physics

Enabling and Supporting Open Science in Heliophysics

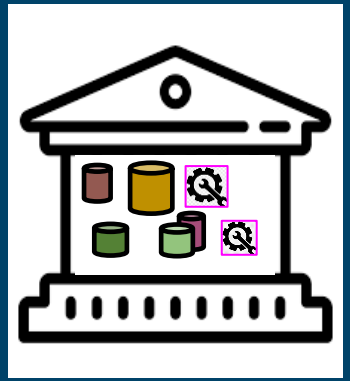
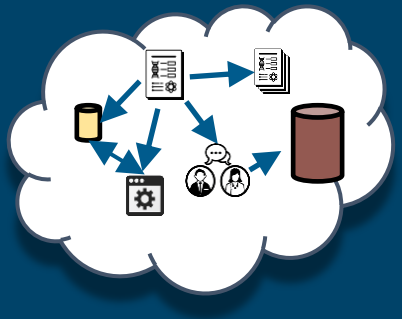
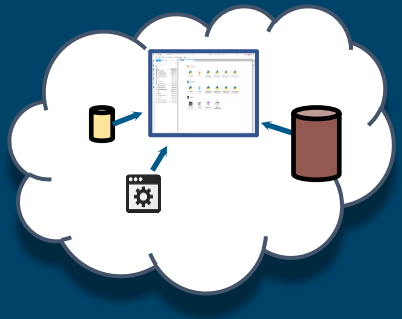
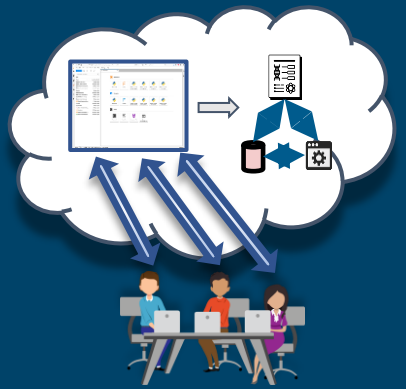
Data
- Clarifying data with unified Data Processing Levels in Heliophysics.
- Leading the way with Data Transparency Levels for researchers.
- Creating DOIs for Solar Physics data.

Metadata
- Adding support for data licenses (SPDX), FAIR vocabularies, and clear citations.
- Interlinking to other resource types (publications, software, containers, etc).
- Improving interoperability with international structures and other sciences (e.g. schema.org).

Services
- Connecting repository API support through helio.data.nasa.gov website.
- Driving collaboration through hosting community computation and data access on HelioCloud.

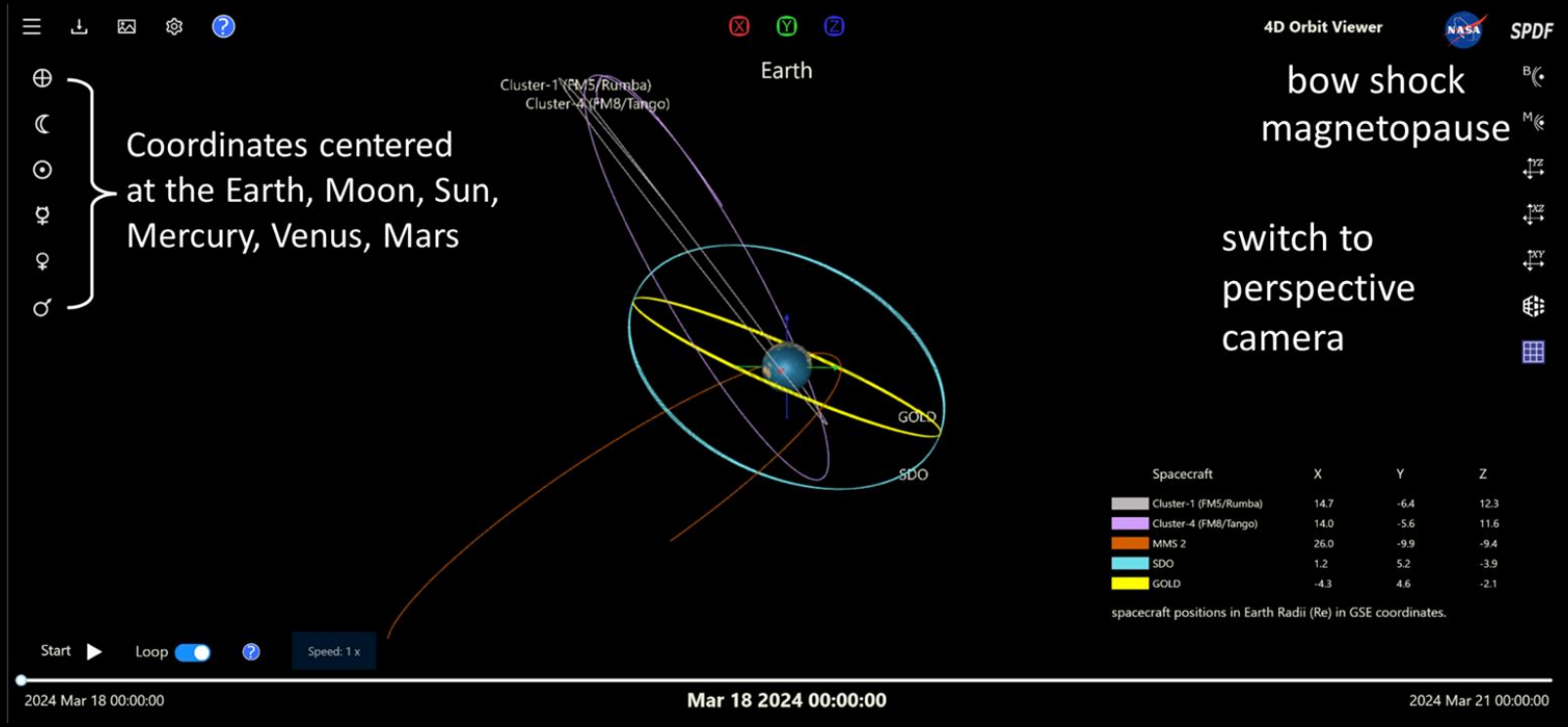
Policy
- Improving flexibility with HDRL Levels of Service.
- Increasing community trust with Desirable Characteristics responses.

Pillars of critical infrastructure : Strategy

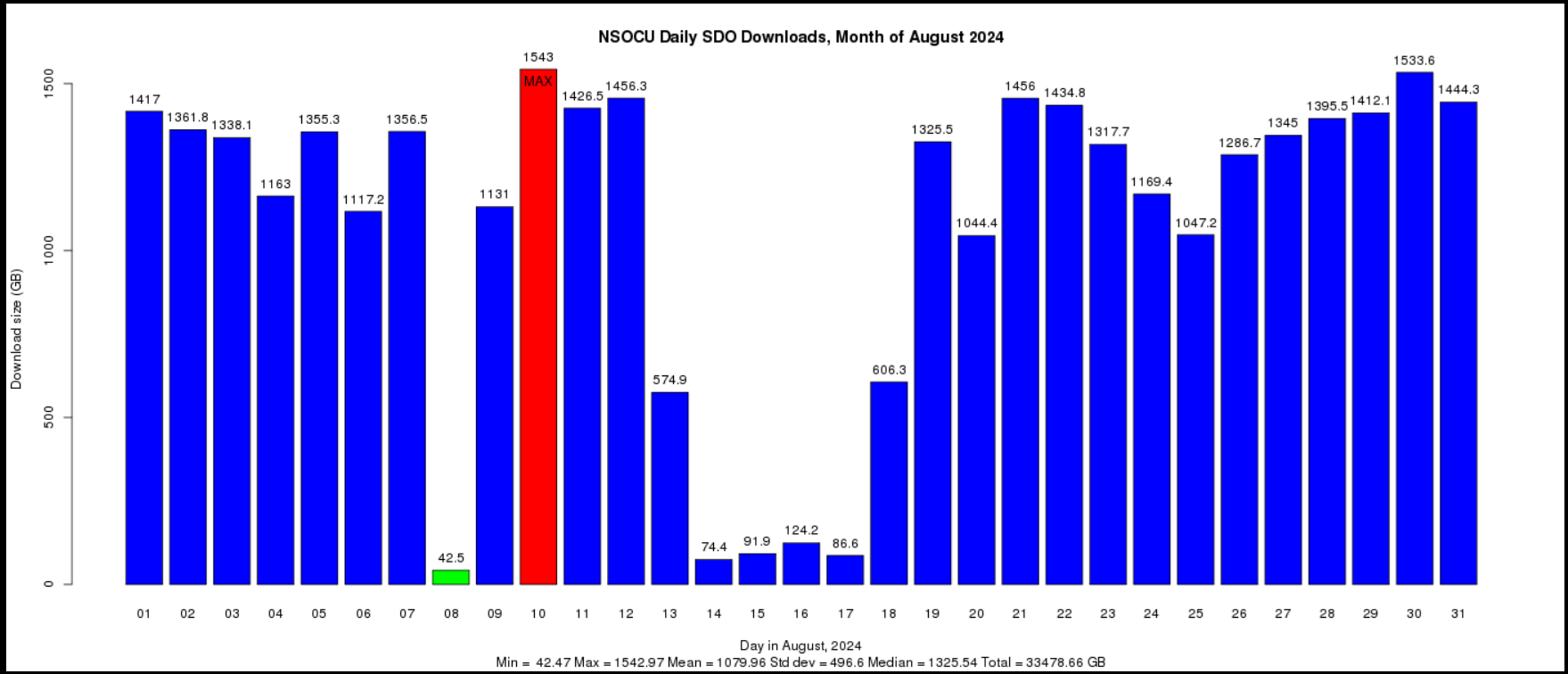
	"Preserve"	*"Discover"*	*"Explore Further"*	*"Extend and Connect"*
Goal	***Provide Foundational Services***	***Enhance Discoverability***	***Unlock Potential***	***Enable Team Open Science***
				
Strategy	*Maintain and upgrade existing archives and services in light of increasing demands driven by Big Data (variety & volume)*	*Increased interlinking of research artifacts, ADS integration, DOIs, improved standards, etc*	*High End Compute close to large (up to ~100 TB) and Big Data (~PB) with software support (AI/ML, PyHC, etc).* Leverage earth science and other platform and expertise.	*Open Science; Collaborative Online Research, Compute, and Publishing Platform & Tools; Open Data; Citizen Science*

4-D Orbit Viewer (160+ Spacecraft)

https://sscweb.gsfc.nasa.gov/4dorbit/



VSO served over 33 TB of SDO data in August 2024



HDRL
HELIOPHYSICS DIGITAL RESOURCE LIBRARY

National Aeronautics and Space Administration

NASA

2024 NASA SCIENCE

NASA Heliophysics Digital Resource Library (HDRL) Update to HPAC

Dr. Jared Bell, Alan Zide, Dr. Susanna Finn, and Dr. Alex Fletcher
Heliophysics Division
October 22, 2024

Heliophysics Data Is a System of Systems

https://doi.org/10.1016/j.asr.2024.01.011

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 73 (2024) 5383–5405

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

Heliophysics Great Observatories and international cooperation in Heliophysics: An orchestrated framework for scientific advancement and discovery

Larry Kepko [a,*], Rumi Nakamura [p], Yoshifumi Saito [t], Angelos Vourlidas [b], Matthew G.G.T. Taylor [o], Cristina H. Mandrini [f], Xóchitl Blanco-Cano [s], Dibyendu Chakrabarty [g], Ioannis A. Daglis [k,w], Clezio Marcos De Nardin [e], Anatoli Petrukovich [u], Minna Palmroth [r,x], George Ho [b], Louise Harra [m], Jonathan Rae [l], Mathew Owens [n], Eric Donovan [i], Benoit Lavraud [c], Geoff Reeves [j], Durgesh Tripathi [h], Nicole Vilmer [q], Junga Hwang [v], Spiro Antiochos [a], Chi Wang [d]

[a] NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[b] Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA

- *Kepko et al.* [2024] describe Heliophysics as a "system of systems."
  - Acoustic Gravity Waves (m) to Corotation Interaction Regions (AU).
- Heliophysics Great Observatory.
  - Mission data.
  - Ground-based observations.
  - Research data and models.
- Calls for long-term, ***open*** curation of heliophysics data and analysis tools.
- This data curation role is central to the Heliophysics Division's vision of enabling cutting-edge science discovery.

The HDRL is critical for HPD to fulfill this role.

HDRL
HELIOPHYSICS DIGITAL RESOURCE LIBRARY

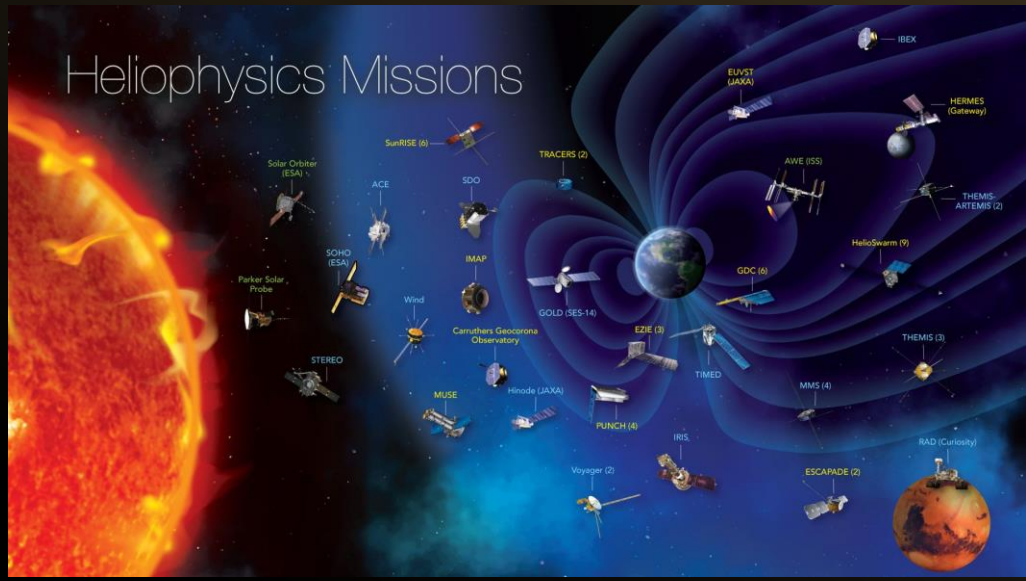
Heliophysics Data is Vast and Increasing!

HDRL role is rapidly growing and evolving.

- Cloud Computing
- Updating Services
- Heliodata website

New demands on HPD/HDRL constantly increasing in both number and complexity.

- Larger mission datasets
- More Heliophysics missions.
- Demands from community producing research data.
  - e.g. ROSES

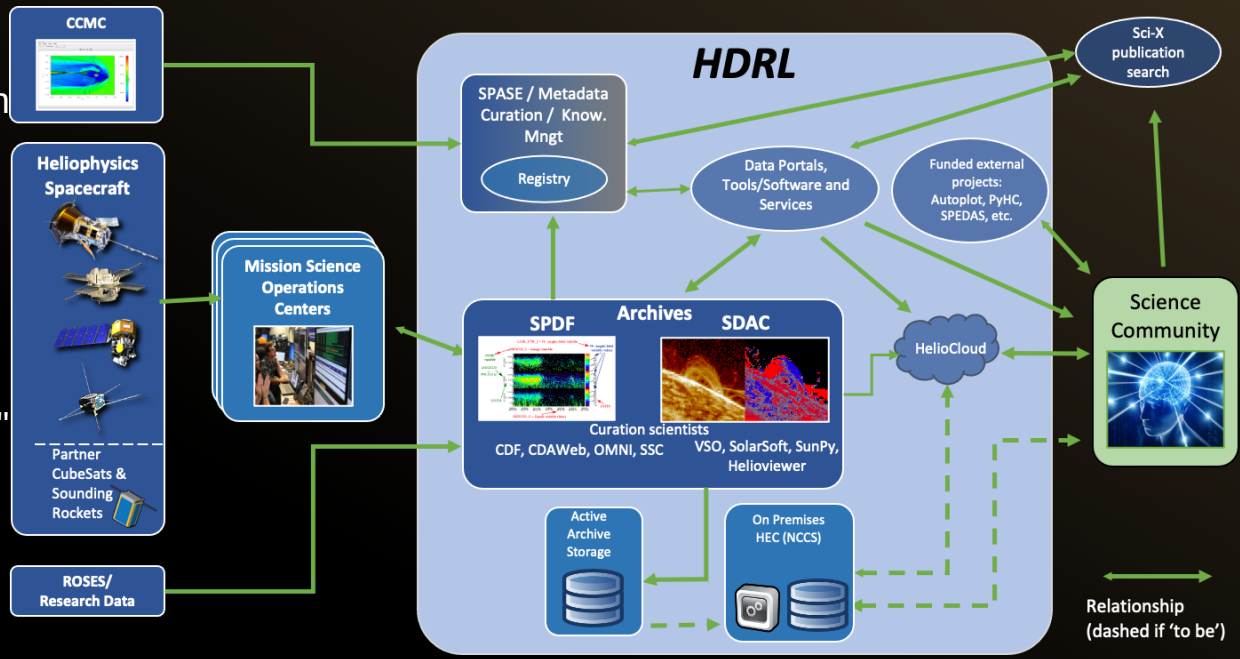


HDRL
HELIOPHYSICS DIGITAL RESOURCE LIBRARY

HDRL Plays a Key Role For The Community

HDRL serves as a key link between spacecraft/research data and the community.

"...embrace and utilize **'big data'** and **ML/AI techniques**... Establish **unified** data formats, metadata, accessibility and discovery via cloud-based computing, and code repositories" (*Kepko et al*. [2024])



HELIOPHYSICS DIGITAL RESOURCE LIBRARY

Key Takeaways from HQ Perspective

HDRL and its services are core to Heliophysics science.

Evolving community needs demand constant and increasing investment.

- New Missions demand more data curation.
- Existing missions require continued curation.
- Legacy missions require updated services and continued curation.

HPD must allocate the resources to provide these core services in a timely manner.

- Community must identify desired data products and services.
- Develop new tools for data analysis and exploration.
- Identify what data products require curation and to what level.
- Maintain data with an Open Science Philosophy (SPD-41a)

HDRL
HELIOPHYSICS DIGITAL RESOURCE LIBRARY

END HQ Portion