Report from the NASA Heliophysics Digital Resource Library

IHDEA Oct 2024

Brian Thomas

POC: Brian Thomas, Project Scientist

Project Scientists: Solar Data Analysis Center (SDAC) – Jack Ireland
Space Physics Data Facility (SPDF) – Robert Candey, Lan Jian
Data & Modeling (HDMC) – Brian Thomas

Project Manager: Tressa Helvey-Kasulke

IHDEA Meeting, Oct 2024, Madrid Spain

HDRL ORGANIZATIONAL CHART

HP Data and Model Consortium / HDMC

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

Registries, Metadata and DOIs for all digital resources; SPASE Data Model.
Heliophysics Data Portal (HDP)
Python and other software integration (PyHC, etc).
Analysis and visualization services ((Py)SPEDAS, Autoplot).
Data upgrades and related services.
HelioCloud initiative with data and software from all groups.

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.

Roles

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

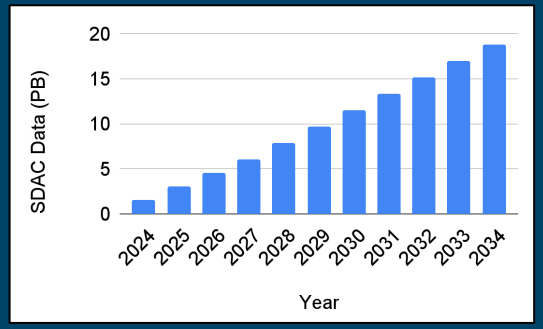
HDRL Holdings

Some statistics (3Q FY24)

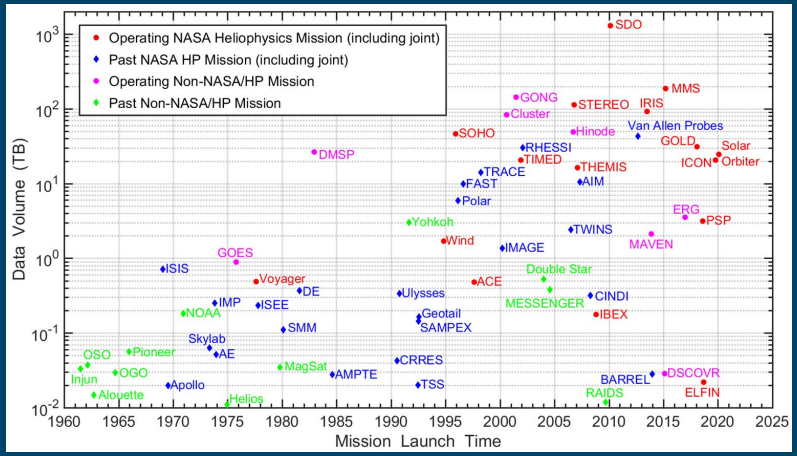
- ~300 million files
- 8000+ datasets
- 2.5+ PB of science data
- 30+ operating missions (& >100 old missions)
- 600+ Instruments
- Over 65 years of data! *(Explorer 3, 1958)*
- *Researcher contributed data*

We have "Big Data"

Volume & Variety

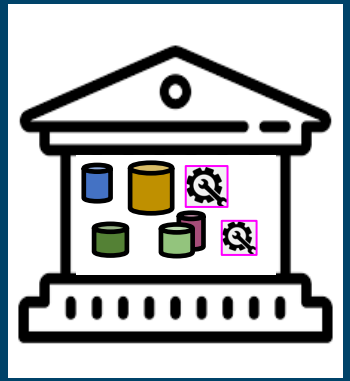
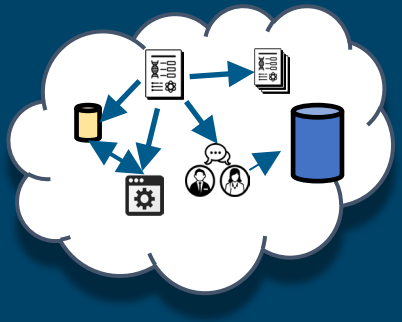
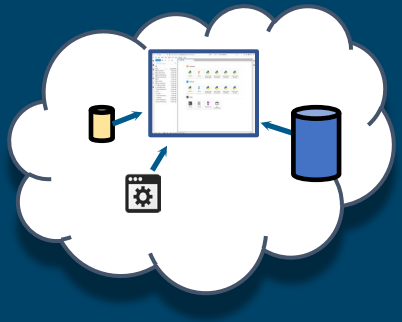
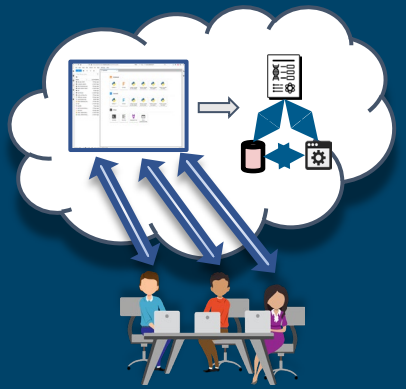


Projected Data Volume Growth to >20 PB



HDRL data holdings span 100+ missions over 65 years.

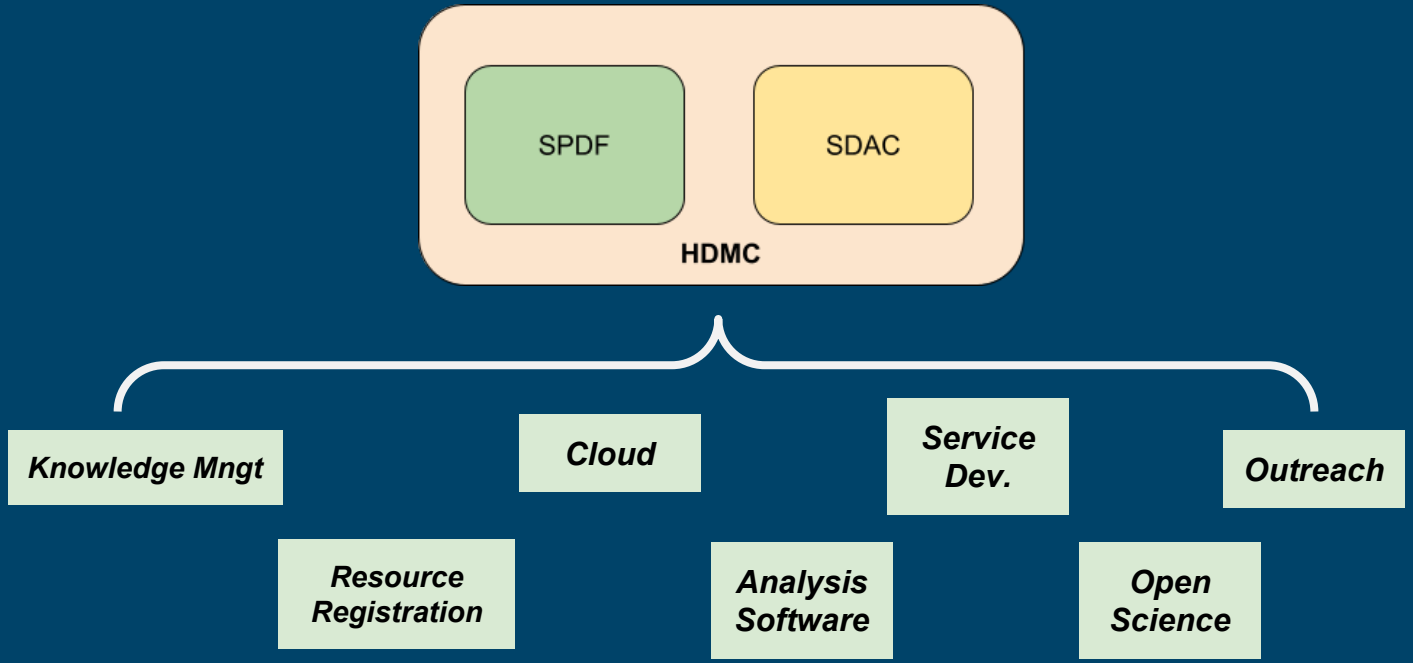
User-Driven Acceleration of Heliophysics Research:
How we get there

	"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*

IHDEA Meeting, Oct 2024, Madrid Spain

HDMC? “HDRL Binder”

HDMC fills gaps in HDRL needs which help unite shared aspects of solar and space physics resources.



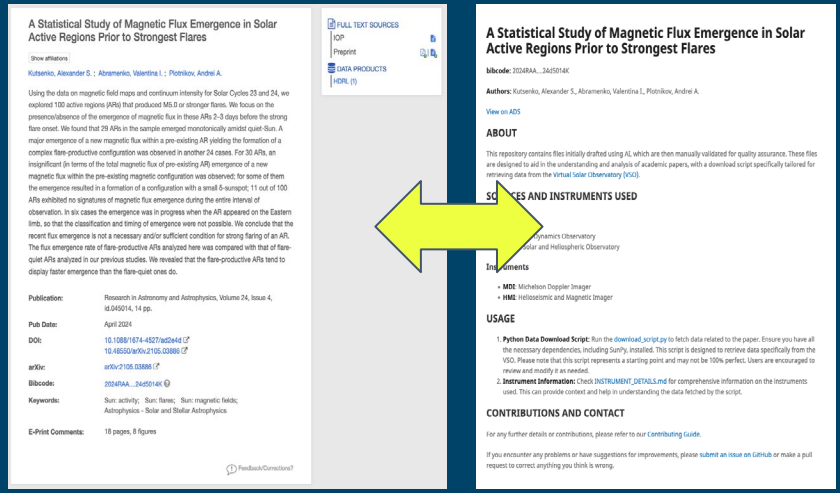
HDMC - Summary of Areas of Work

Knowledge Management	**Foundational** **Discovery**	Standards and Data model development, Resource Tagging Automation & Tool development, OSCDO/OMB Compliance, Semantics, Data Science
Resource Registration	**Foundational**	DOI / data registration for discovery, Metadata curation. OSCDO/OMB Compliance
Cloud	**Explore Further**	HelioCloud and Cloud infrastructure for HDRL (grant), Compliance to OCSDO/OCIO, Cloud operations and infrastructure support
Service Development	**Discovery** **Extend & Connect**	Services to enable system science; Helio.data and Space Physics HEK (grant)
Software	**Foundational,** **Explore Further**	Providing means for community to analyze heliophysics data (ex SPEDAS, HAPI, PyHC; grants)
Outreach	**Foundational**	Understanding community and communicating back, compliance with NASA (comms / HQ).
Open Science	**Extend & Connect**	Addressing OCSDO/OMB requirements, prototyping services, policy development

Past Year Highlights: Discover

SciEx (ADS) collaboration

- First interlinked papers for heliophysics data with ADS using Machine Learning (ML)
- UAT modifications for Space Physics
- Added Space Physics journals



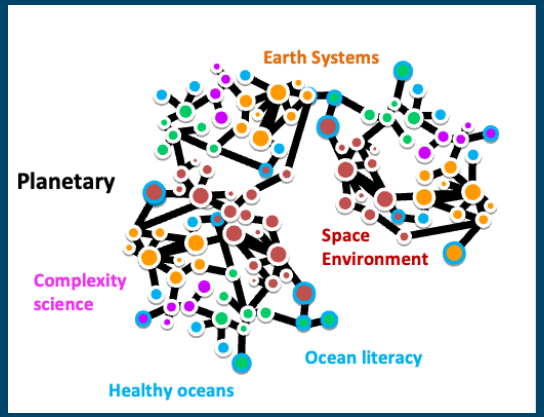
***Interlinked Papers & Data**: ADS/HDRL dataset to paper interlinking w/ ML*

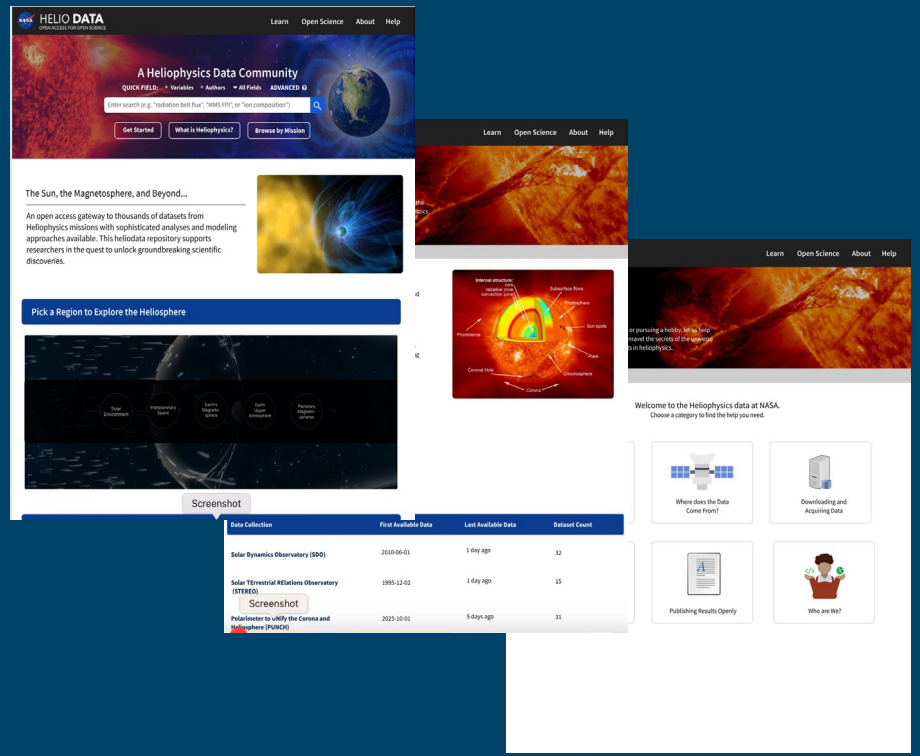
UNIFIED ASTRONOMY THESAURUS
a community supported, open source project from the American Astronomical Society

***Improved Space Physics Search**: Collaboration with Sci-X to add space physics terms to the UAT (used by Sci-X)*

Past Year Highlights: Discover

- **Integration with ODIS**
  - SPASE data discoverable by Oceanographers. (late 2024)
- **Helio.data.nasa.gov website**
  - Public release in 2025
  - Based on SPASE, targeting early researchers





***Integration with Earth Science**: uploaded SPASE records to ODIS. Space weather data at HDRL soon discoverable by oceanographers (late 2024).*

***New Website**: Early researcher target audience. SPASE-based. (Release in 2025)*

Past Year Highlights: Explore Further

Cloud research environment for Heliophysics ("HelioCloud")

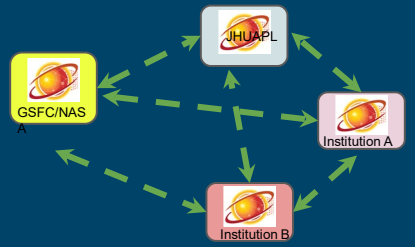
- Participation in IHDEA working group for distributed software environ.
- CloudCatalog released
  - Python-based search and retrieval of data.
- Growth in participation
  - 30+ Research Science Projects (past year)
  - 10+ published research papers (past year)
  - 160+ registered Users, >~ 50 active users (activity in past month)
- Uploaded ~> 1 PB of science data to S3
  - SuperMAG added
  - High-level HDRL missions
  - ML contributed datasets (FDL)
- Supported PyHC, Heliophysics MUREP Interns, HDRL Outreach, Nepali student engagement events.

Cloud Infrastructure ("HSDCloud")

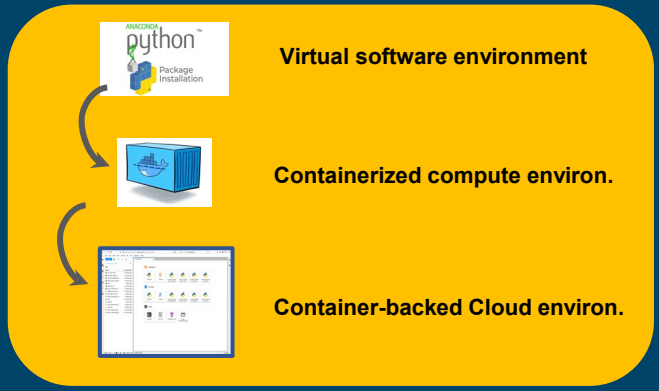
- 10+ hosted projects
- Example: Eclipse viewer

On Premise Compute (GSFC High End Compute Environment)

- New hire in FY25
- Updating shared software environment (HelioCloud)



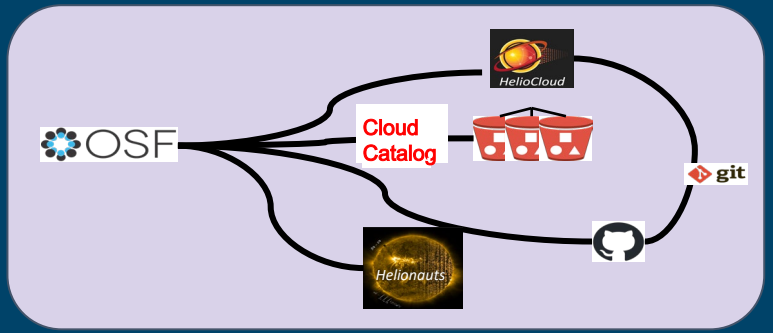
HelioCloud CloudCatalog: *Facilitates cloud-hosted data discovery and access. Based on HAPI.*

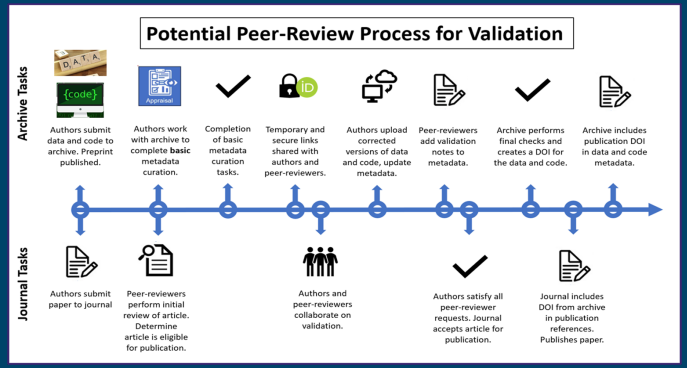


Shared Environments: *Define various environments which build on prior and may be used to share software, enhance Open Science publication and shared research environments.*

Past Year Highlights: Extend and Connect

- **Open Science**
  - FAIR assessment
  - Levels of Service
  - Prototyping services, workflows for Open Science
    - Zenodo, OSF-based prototypes
- **Outreach**
  - AGU, TESS, AMS, Shine, GEM, CEDAR
  - Workshops
  - Analysis Working Groups (in Development)
- **Common Python-based Heliophysics Data Analysis Environment**
  - Continued development to add GPU/ML acceleration.
  - Collaboration with GSFC HEC, PyHC, IHDEA and 2i2c communities.





Open Science Services: *Work to prototype open science services (above) and workflows (below) for heliophysics. Identification of gaps in infrastructure (ongoing).*

Outreach Indicates Critical Needs

Outreach to community (‘20 questions’, workshops) indicates:

- **Better access for big data** (10+ TB; ex. MMS, SDO, etc)

- **Improved discovery** across all domains of heliophysics
  - **What is related to what?**
    - Interlinked research artifacts. Find data used by paper.
  - **Researcher-facing Search**
    - by Region, Phenomena, Methodology/Models
  - **“Getting started” is (too) hard**
    - Vast landscape of resources is hard to understand without help

- **Open Science is “too hard”**. Need for infrastructural support.

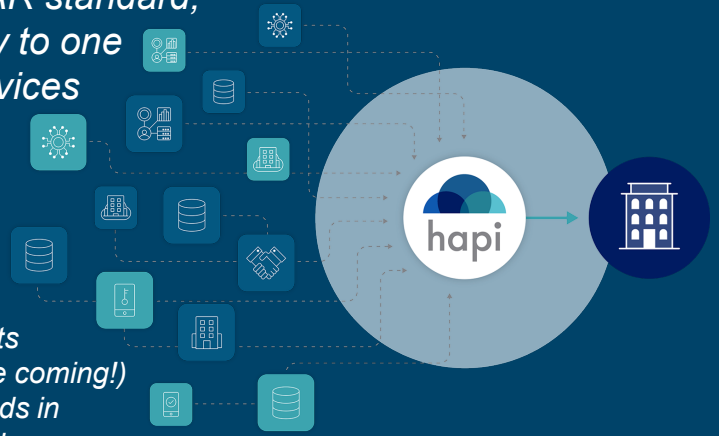
Thank you!

Backup Slides

Software Supporting Analysis

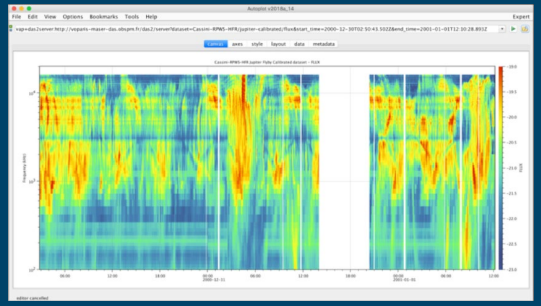
HDMC supports research enabling software for the heliophysics community and beyond!

***HAPI**, a COSPAR standard, provides a many to one interface for services & data. Used widely & internationally.*



- *~10,000 datasets (and many more coming!)*
- *9 access methods in*
- *7 programming languages*
- *9 HAPI Servers in Helio (HDRL, CCMC, and non-NASA)*
- *More coming soon (e.g. SWPC, Madrigal and even Earth Sci)!*
- *https://hapi-server.org/servers/*

***SPEDAS / pySPEDAS** is a leading analysis environment for space physics mission data.*



***Autoplot** visualization software*

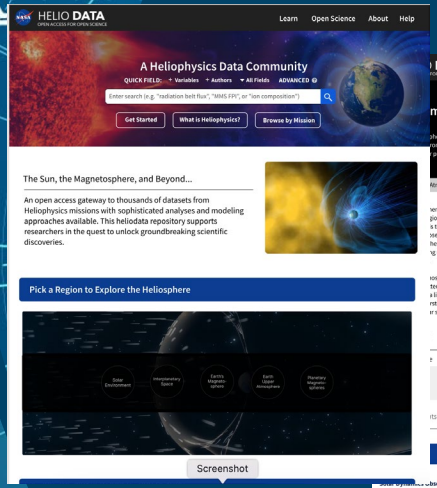
Software Supporting Analysis: Community

- **PyHC** : Python in Heliophysics Community
  - https://pyhc.org
  - > 80 projects, 7 core
    - pySPEDAS, spacepy, sunpy, plasmapy, hapi, pysat, Kamodo
  - Community of >100 Python software developers
  - Bi-annual Python in Heliophysics summer school with 424 attendees for 2024 (~300 for 2022).
  - Common software environment (in Python)
  - Development of Python software standards in Heliophysics https://github.com/heliophysicsPy/standards/pulls

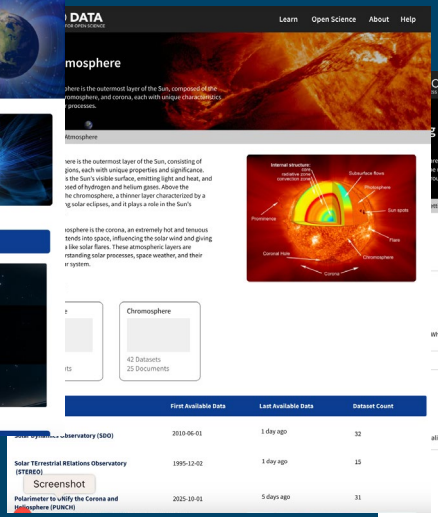
Service Development

Helio.data Website and API

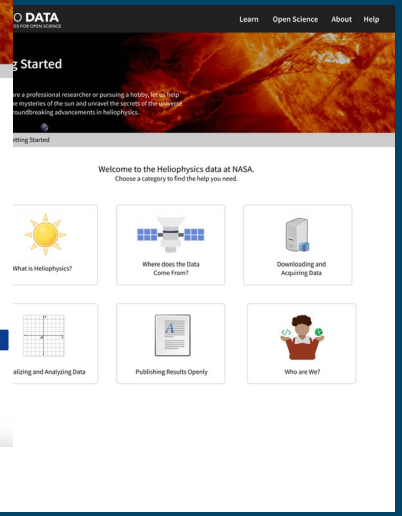
HDRL resources discoverable from researcher point of view, better support early researchers and public



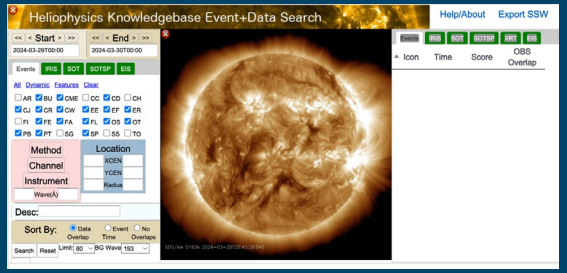
Top Page
search by keyword, region browsing



Example Region Page
Interlinked resources, docs



Getting Started Page
Interlinked resources, docs, how to publish open science



Heliophysics Event Knowledgebase

Extension to add space physics events in HEK (w/ SDAC)

Outreach

User outreach to understand need, improve services

- 20 questions exercise
- Infrastructure Workshops
- User engagement sessions at SHINE, CEDAR, AGU 2024
- Leading planning of NASA Software for the SMD Workshop with OCSDO
- TESS and AGU 2024 sessions on Open Science, Data Env. Townhall (+NSF and NOAA)
- Developing Analysis Working Groups in Heliophysics

Community Building / Meeting Support

- HQ Heliophysics Big Year effort (under discussion)
  - CDAW-like workshops
  - ISTPNext 2024/2025
  - Space Apps Challenge 2024
  - ISWAT 2025
- Supporting many Heliophysics workshops with HelioCloud
- DASH (Data Analytics and Software in Heliophysics)
- IHDEA (International Heliophysics Data Environment Alliance)
- IHDWG (Inter-agency Heliophysics Data Working Group)

Outreach Indicates Critical Needs

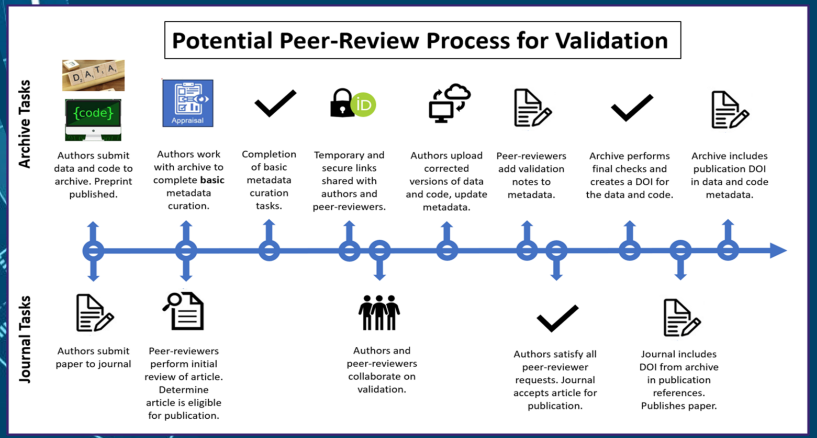
Outreach to community (‘20 questions’, workshops) indicates:

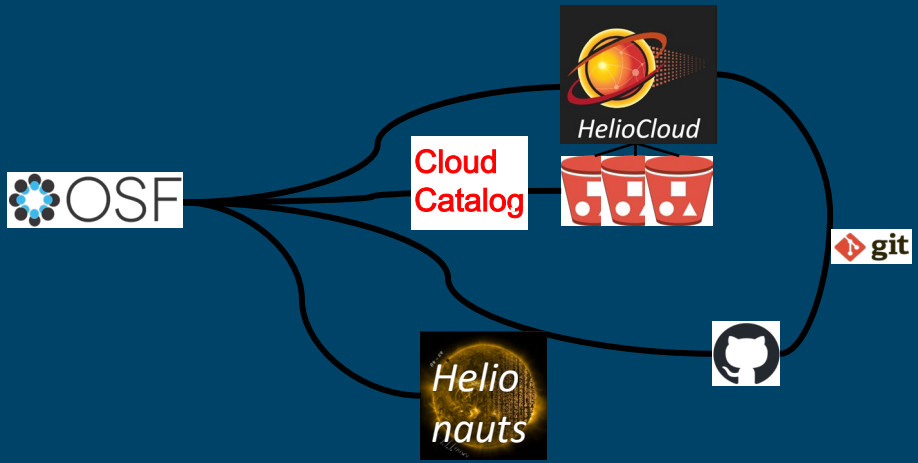
- **Better access for big data** (ex. MMS, SDO)
- **Improved discovery** across all domains of heliophysics
  - **What is related to what?**
    - Interlinked research artifacts. Find data used by paper.
  - **Researcher-facing Search**
    - by Phenomena, Region, Methodology
  - **“Getting started” is (too) hard**
    - Vast landscape of resources is hard to understand without help
- **Open Science is “too hard”**. Need for infrastructural support.

Making Open Science *Easy* in Heliophysics

Strategy: Determine the infrastructure we need at HDRL by

- **Evaluation of HDRL** services and data for compliance to FAIR, SPD-41a / OSCDO & OPM compliance
- **Increase the Flexibility** of HDRL services to accommodate a spectrum of openness and increase FAIR
- **"Practice" open science** projects to find gaps and barriers in processes, services and standards
- **Partnering** with OSCDO, Publishers, ADS, Zenodo, & others to determine potential processes
- **Community Involvement and Outreach** via workshops and SMD and international WGs

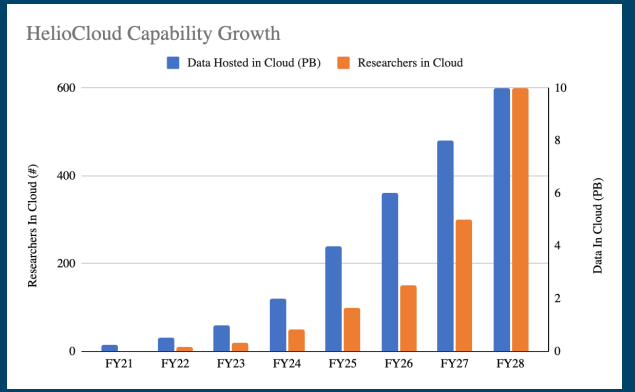




OLD SLIDES

FY24: Strategy

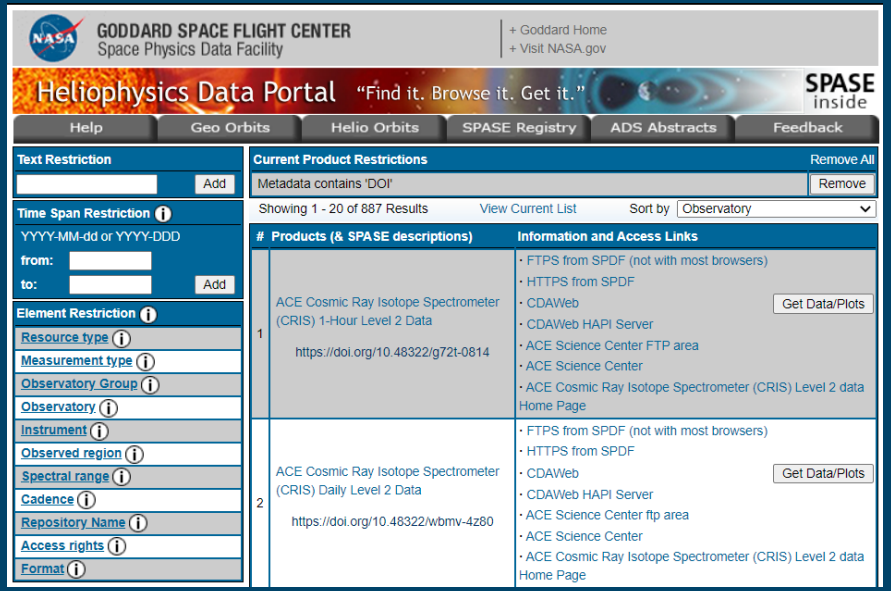
- **Discovery service improvements**
  Enhanced discovery site/service. Enhanced event list and API support.
- **Deploy enhanced compute capabilities**
  HelioCloud, NCCS into production
- **Increase data available for analysis, discovery**
  In cloud and on premise to petabyte volumes, push for SPASE evolution, CCMC/HDRL prototype
- **Engagement**
  IHDEA / Workshops : to engage community / develop partnerships and shared vision / standards (SPASE, STC, ??) / collaboration (cloud, registry) / DASH 2?
- **Open Science / FAIR**
  Outreach, definition of strategy, projects, workshop(s), IHDEA wg?
- **Conduct outreach to engage broader community**
  Spread word and get broader adoption, engagement; user group
- **Deploy new infrastructure**
  HelioCloud service components, other service API(s)



HelioCloud: *Planned growth in researchers and data hosted by the instance at NASA*

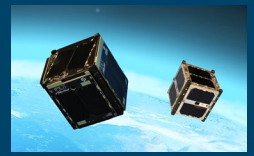
FY24(+): Discover

- **HDP** — Improved user interface. Standardization of API. Search by phenomena.
- **HEK** — Ingestion / full support for non-solar event data.
- **ADS** — Integration of software, other research artifacts to interlink with NASA datasets and Heliophysics publications
- **Improved support for Event Lists**
- **HelioCloud -** develop science database, data registry searches across instances.
- **CDAWeb data browser -** add interactive search and display.



FY24(+): Preserve

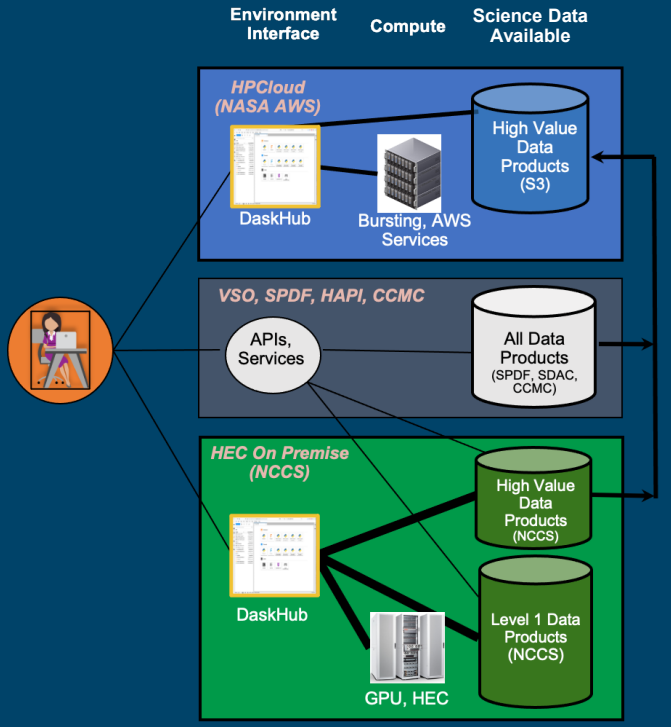
- **Enhanced data ingest and registration**
  - Develop more ***streamlined registration*** (DOIs)
  - Develop ***improvements to metadata standards*** to support requirements
    - Better support registration of software
    - Better support for open science publishers
  - Support data from smallsats, balloons, rockets, ground-based observatories.
- **Transfer of Multi-PetaByte Data to SDAC**
  - SDO Level 0 data (~6PB) from SDO JSOC to SDAC.
- **Offline backup of all data currently at SDAC**



JSOC
SDP

FY24(+): Explore Further

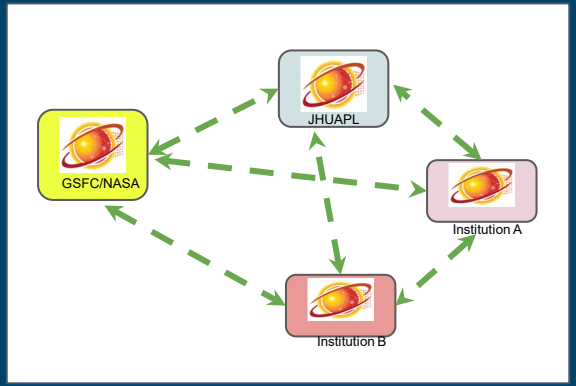
- **HelioCloud**
  - Release easily deployed cloud-based deployable environment (Blue box), AGU 2023
  - Ingest up to 8 PB of data to cloud
- **On premise research environment for Big Data**
  - In production at NCCS (Green box)
  - Implementation of DaskHub environment
- **Transfer of AIA Level 1 data from SDO JSOC to NCCS**
  - Scientific use of AIA Level 1 data on NCCS computational environment using HelioCloud shared data analysis environment and DaskHub platform (deployed at NCCS)
- **Helioviewer**
  - Add capability to show selected content from CCMC and SPDF
- **Support AI/ML-ready data, Model / Data comparison**
  - With priority for cloud-based analysis by public, CCMC collaboration



New Compute Environments: *In operation in the Cloud (Blue) and @ NCCS (Green) with common software and platform (DaskHub). Thickness of lines indicate throughput.*

FY24(+): Extend and Connect

- **HelioCloud**
  - Provide easily deployed framework for deployment at other institutions
  - Add tools to add researchers to publish Open Science
  - Add support for sharing notebooks, containerized software and other research artifacts
  - Develop citizen science projects hosted in HelioCloud
- **Open Science Software Publication**
  - Shared software environment for Heliophysics research (Python)
  - Community engagement to develop best practices, standards for software publication



HelioCloud: *Research artifacts may be easily 'published' and shared with others at other institutions.*

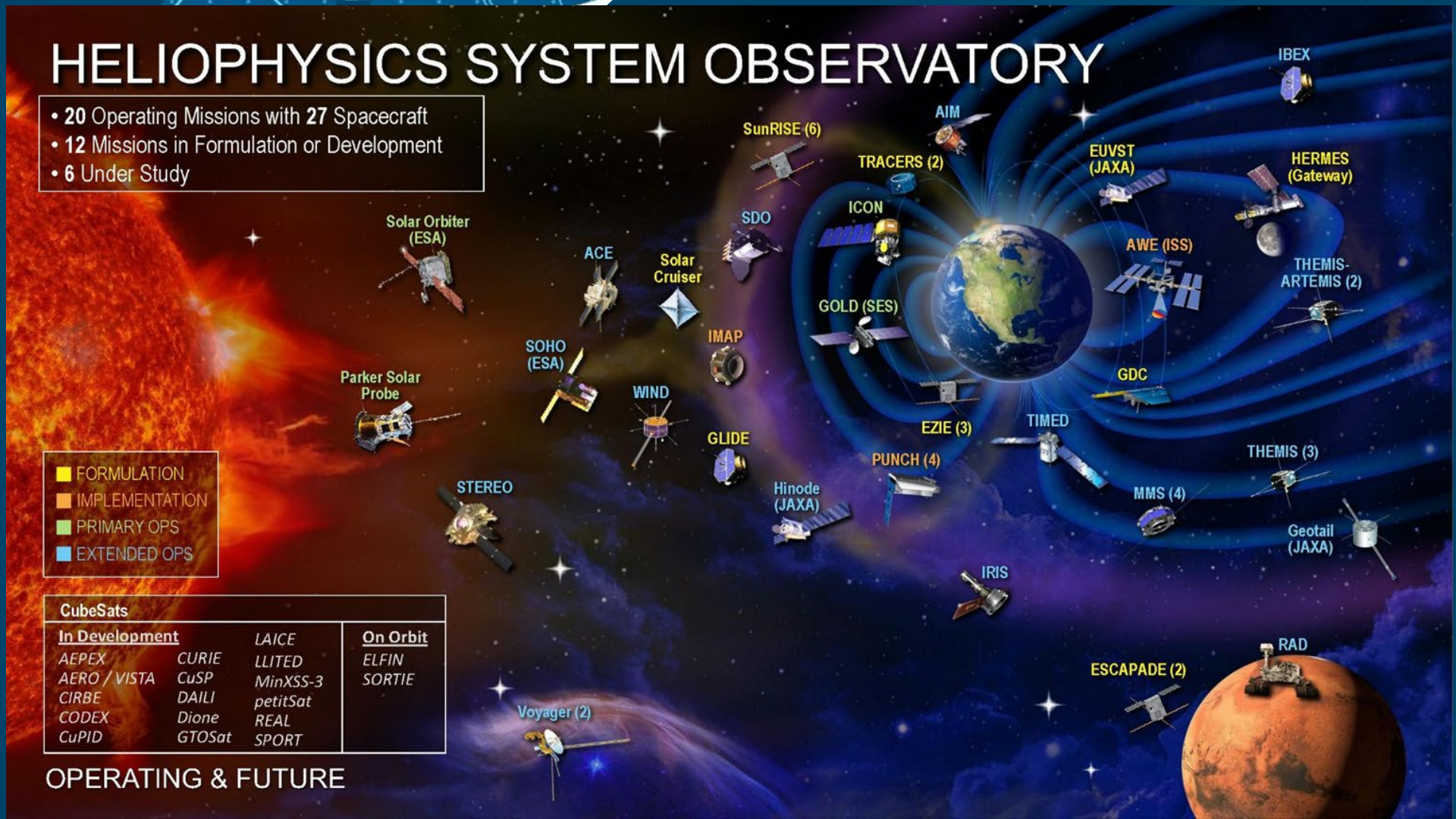
New FY24 Initiatives (Summary)

- Open Science (OS)
  - FAIR assessment of services
  - Open Science software workshop
  - Refine understanding of infrastructure needed for OS
- Outreach
- Discovery
  - helio.data.nasa.gov
  - HEK / Space Physics
  - ADS / SDE collaboration

BACK UP SLIDES

HELIOPHYSICS SYSTEM OBSERVATORY

- **20** Operating Missions with **27** Spacecraft
- **12** Missions in Formulation or Development
- **6** Under Study



- FORMULATION
- IMPLEMENTATION
- PRIMARY OPS
- EXTENDED OPS

CubeSats

In Development			On Orbit
AEPEX	*CURIE*	*LAICE*	*ELFIN*
AERO / VISTA	*CuSP*	*LLITED*	*SORTIE*
CIRBE	*DAILI*	*MinXSS-3*	
CODEX	*Dione*	*petitSat*	
CuPID	*GTOSat*	*REAL*	
		SPORT	

OPERATING & FUTURE

Vision

Where the System Observatory Comes Together

The HDRL enables the scientific analysis goals of the Heliophysics System Observatory:

- ***Provisioning and curation of scientific big data*** from many sources, PB volumes; (the Foundation: data, metadata, standards)
- ***Support for data analysis and modeling*** in multiple computational environments;
- The design and implementation of a ***collaborative open science infrastructure.***

Individual missions can do great science

Unlocking groundbreaking systems science requires the HDRL