IMPACT
Interagency Implementation
and Advanced Concepts Team

Science Discovery Eng

Enabling Search and Discovery of NASA’s Data and Inf

March 22, 2023

Kaylin Bugbee
Science Discovery Engine Project Lead | Interagency Implementation and Advanced Concepts Team (IMPACT)
Applications Data Manager | Earth Science Branch (ST11)
Marshall Space Flight Center/NASA
Huntsville, Alabama 35812, USA

Outline

1 Overview of the Science Discovery Engine

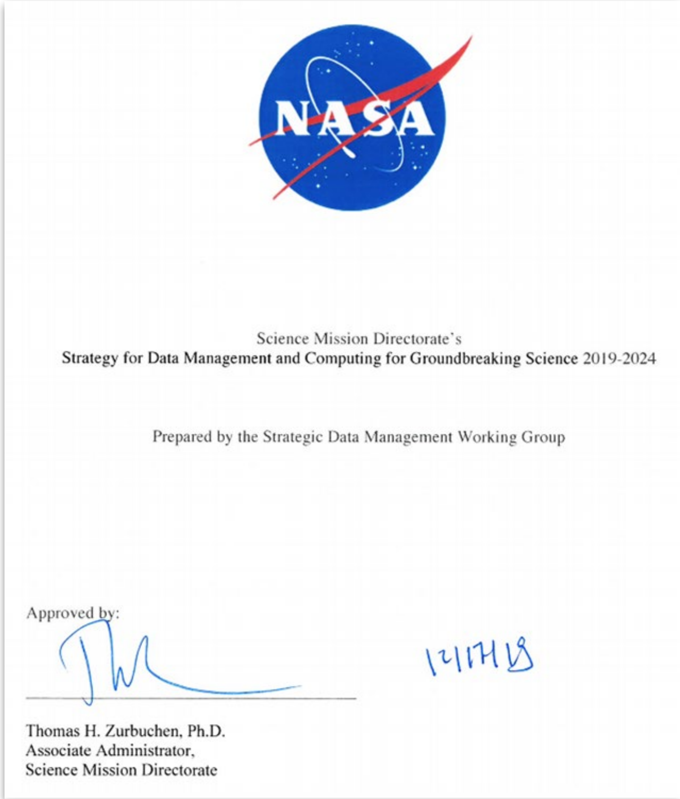
2 Alignment with Workshop Goals

3 Implement AI/ML Into Workflows

4 Enable Analyses for New Science Discoveries

5 Discussion

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



Science Mission Directorate's
Strategy for Data Management and Computing for Groundbreaking Science 2019-2024

Prepared by the Strategic Data Management Working Group

Approved by:

12/17/18

Thomas H. Zurbuchen, Ph.D.
Associate Administrator,
Science Mission Directorate

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** .
- **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

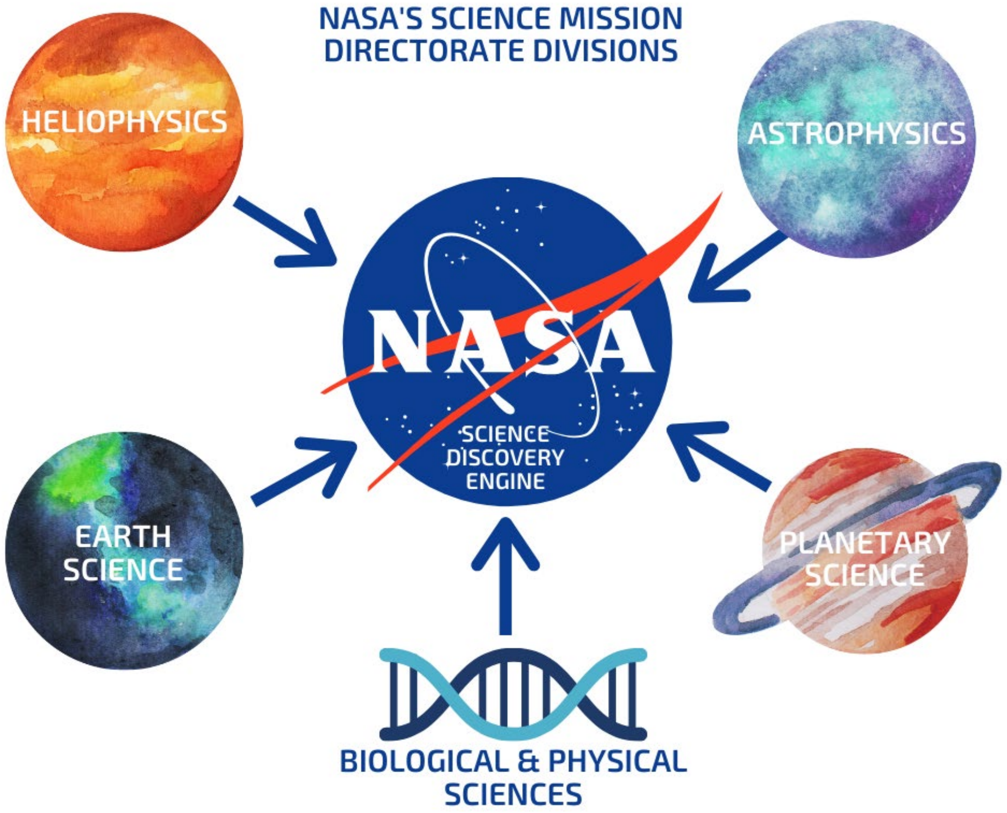
Goal 3: Harness the Community and Strategic Partnerships for Innovation

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

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	
1.1	Develop and implement a **consistent open data and software policy** tailored for SMD	2.1	Establish **standardized approaches for all new missions** and sponsored research that encourage the adoption of advanced techniques	3.1	Develop **community of practice and standards group**
1.2	Upgrade capabilities at existing archives to **support machine readable data access using open formats and data services**	2.2	Integrate investment decisions in High-End Computing with the strategic needs of the research communities	3.2	Partner with **academic, commercial, governmental and international organizations**
1.3	Develop and implement a SMD data catalog to support discovery and access to complex scientific data across divisions	2.3	Invest in capabilities to use commercial cloud environments for open science	3.3	Promote opportunities for continuous learning as the field evolves through collaboration
1.4	Increase transparency into how science data are being used through a free and open unified journal server	2.4	Invest in the tools and training necessary to enable breakthrough science through application of AI/ML		

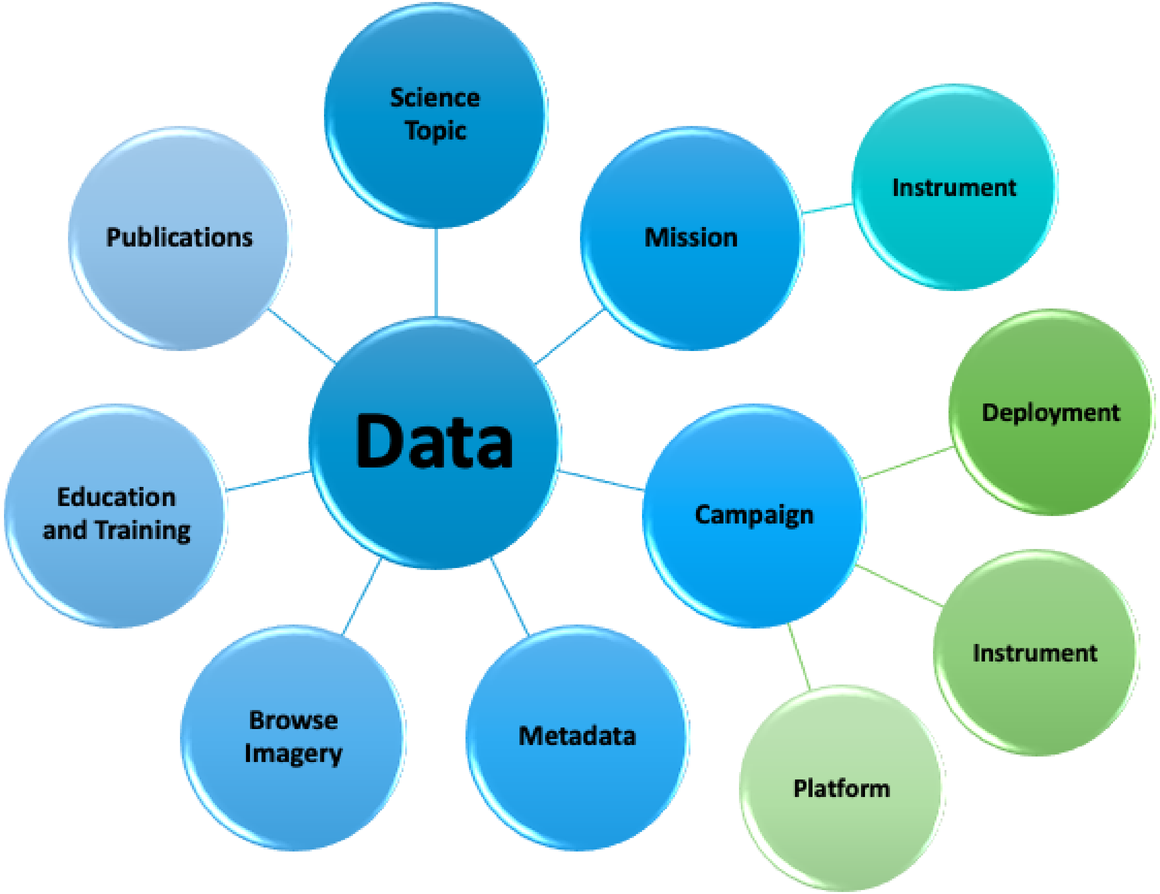
SMD Science Discovery Engine (SDE) Goals

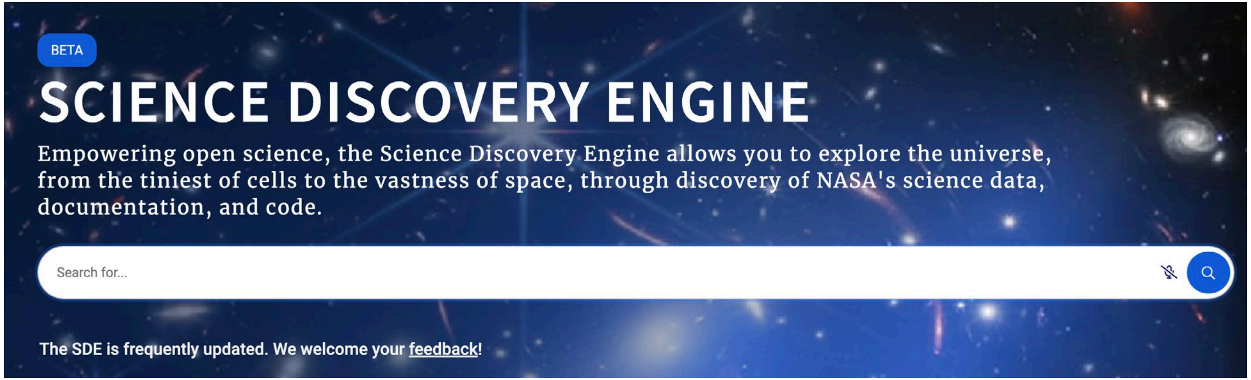
- Enable discovery of open science objects including
  - Data, software, models, images and documentation
- Beta SDE released in December 2022
  - Curation & inclusion of additional content is ongoing
  - New features added quarterly



Use Case

Enable the discovery of SMD **data, software, code, publications, models, documentation** and other relevant information in context to make research more efficient.



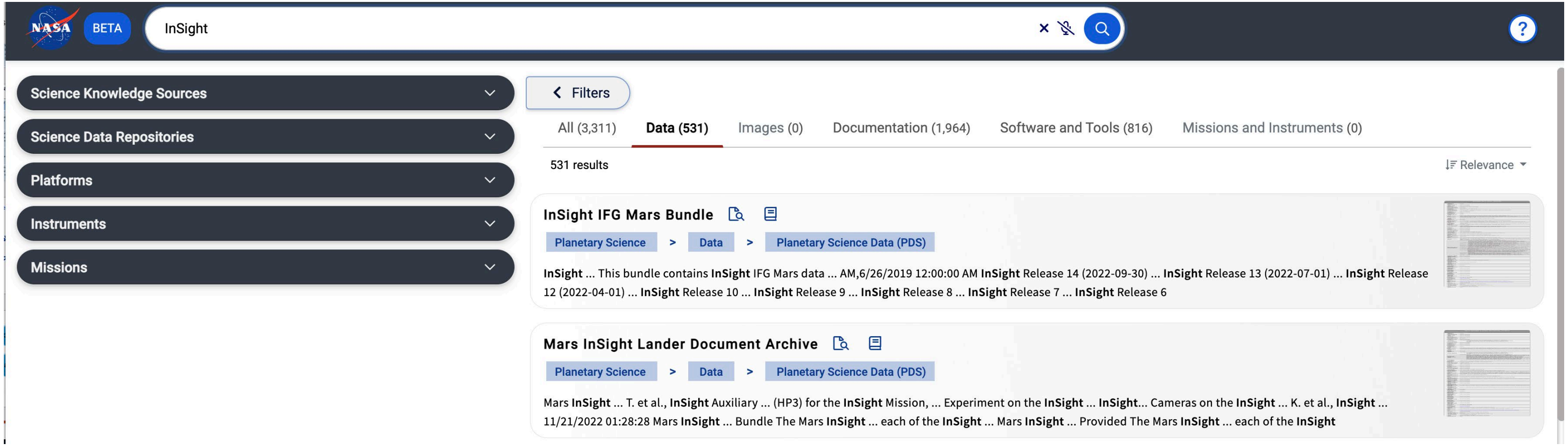


BETA

SCIENCE DISCOVERY ENGINE

Empowering open science, the Science Discovery Engine allows you to explore the universe, from the tiniest of cells to the vastness of space, through discovery of NASA's science data, documentation, and code.

The SDE is frequently updated. We welcome your feedback!



531 results

InSight IFG Mars Bundle

Planetary Science > Data > Planetary Science Data (PDS)

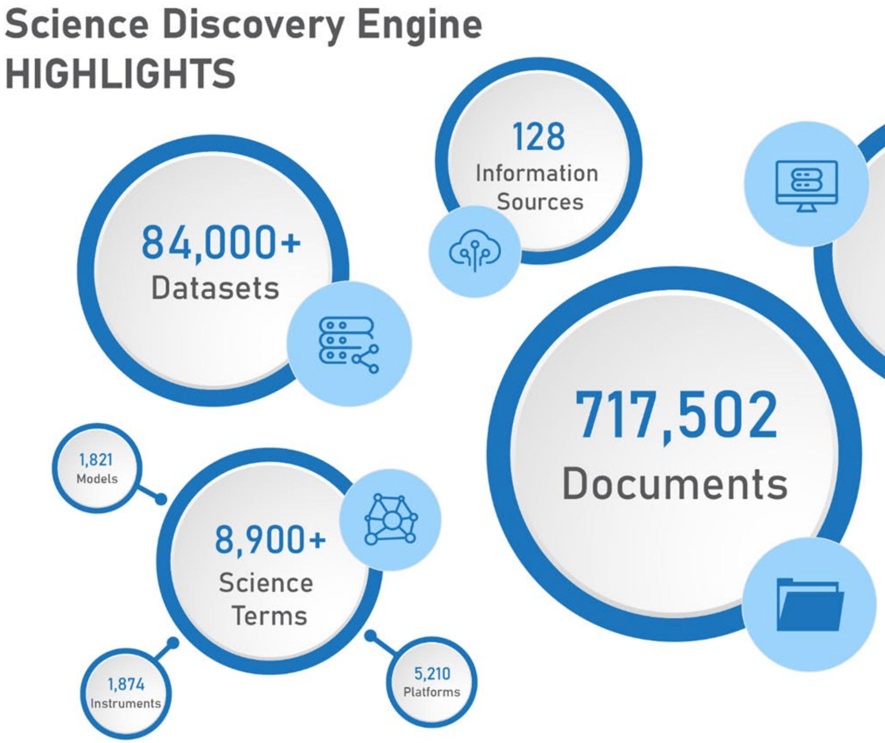
InSight ... This bundle contains **InSight** IFG Mars data ... AM,6/26/2019 12:00:00 AM **InSight** Release 14 (2022-09-30) ... **InSight** Release 13 (2022-07-01) ... **InSight** Release 12 (2022-04-01) ... **InSight** Release 10 ... **InSight** Release 9 ... **InSight** Release 8 ... **InSight** Release 7 ... **InSight** Release 6

Mars InSight Lander Document Archive

Planetary Science > Data > Planetary Science Data (PDS)

Mars **InSight** ... T. et al., **InSight** Auxiliary ... (HP3) for the **InSight** Mission, ... Experiment on the **InSight** ... **InSight**... Cameras on the **InSight** ... K. et al., **InSight** ... 11/21/2022 01:28:28 Mars **InSight** ... Bundle The Mars **InSight** ... each of the **InSight** ... Mars **InSight** ... Provided The Mars **InSight** ... each of the **InSight**

SDE Beta Highlights



Science Discovery Engine
HIGHLIGHTS

84,000+ Datasets

128 Information Sources

44,723 Software Tools Models

717,502 Documents

8,900+ Science Terms

1,821 Models

1,874 Instruments

5,210 Platforms

SDE Alignment with Workshop Goals

"Embracing new technologies to implement AI/ML in their workflows to advance their analyses for new science discoveries"

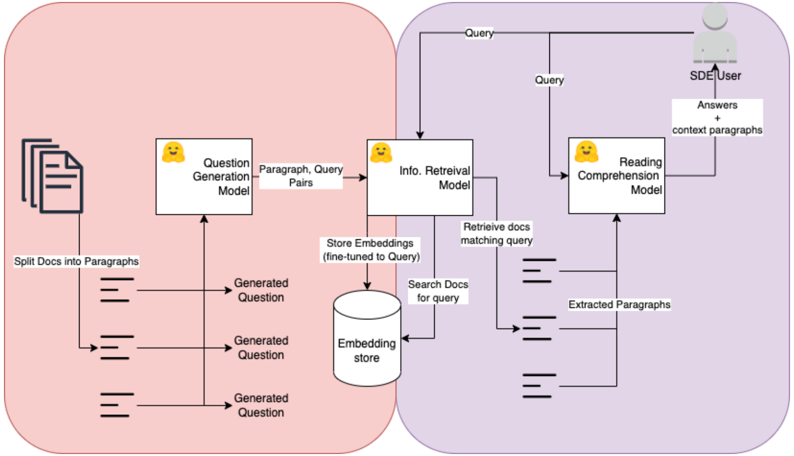
Implementing AI/ML into SDE Workflows

- We can leverage LLMs to automate curation workflows into the SDE. This could include:
    - Thematic area content curation
        - Automatically classify documents into thematic areas such as ‘climate change,’ ‘space weather’ etc
    - Content type curation
        - Automatically classify documents into appropriate category types such as data, images, documentation, software, models etc.
    - Acronym / Word Disambiguation
        - Disambiguate acronyms or words based on context
    - Entity Recognition

Implementing AI/ML into SDE Workflows

We can increase the relevancy of results in the SDE using question answering and dense information retrieval with language models (Neural Search)

- **Dense IR models encode the query and the document collection into a vector space**, where the similarity between the query and each document can be computed. The resulting dense vectors can be used to retrieve a set of candidate documents that are most likely to be relevant to the query. This approach is effective for complex queries and long documents, where traditional retrieval methods may struggle to capture the full meaning and context of the text.
- **Question-Answering,** on the other hand, is a task that aims to provide direct and concise answers to natural language questions based on retrieved / relevant documents
- Performance of the models depend on underlying transformer model and will be improved by pre-training on domain specific corpus

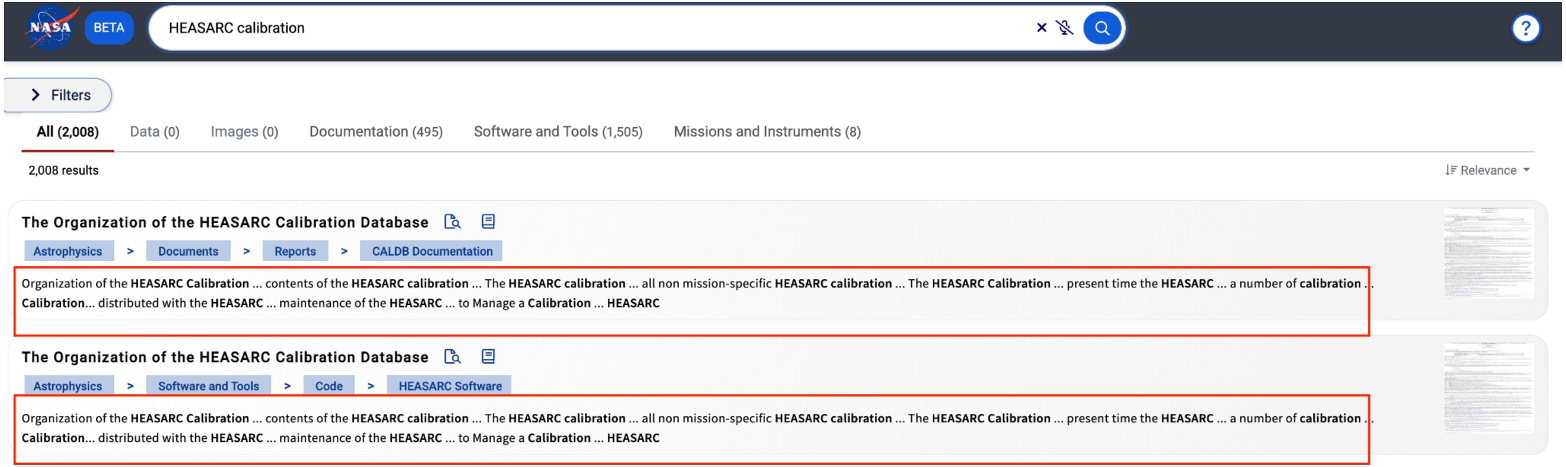


Dense Retrieval and Question Answering

Implementing AI/ML into SDE Workflows

We can increase the relevancy of results in the SDE using document/web page summarization

On the results page, language models can be used for summarizing documents and webpages, which can help user in identifying the relevant content faster.


BETA
HEASARC calibration
Filters
All (2,008)
Data (0)
Images (0)
Documentation (495)
Software and Tools (1,505)
Missions and Instruments (8)
2,008 results
Relevance
Astrophysics > Documents > Reports > CALDB Documentation
Astrophysics > Software and Tools > Code > HEASARC Software

The Organization of the HEASARC Calibration Database

Organization of the **HEASARC Calibration** ... contents of the **HEASARC calibration** ... The **HEASARC calibration** ... all non mission-specific **HEASARC calibration** ... The **HEASARC Calibration** ... present time the **HEASARC** ... a number of **calibration** .. **Calibration**... distributed with the **HEASARC** ... maintenance of the **HEASARC** ... to Manage a **Calibration** ... **HEASARC**

The Organization of the HEASARC Calibration Database

Organization of the **HEASARC Calibration** ... contents of the **HEASARC calibration** ... The **HEASARC calibration** ... all non mission-specific **HEASARC calibration** ... The **HEASARC Calibration** ... present time the **HEASARC** ... a number of **calibration** .. **Calibration**... distributed with the **HEASARC** ... maintenance of the **HEASARC** ... to Manage a **Calibration** ... **HEASARC**

Advancing Analyses for New Science Discoveries

The SDE can accelerate analyses for new science discoveries by allowing users to discover SMD AI/ML

- Models
- Datasets
- Documentation

In addition to all of the other relevant data and information available across SMD.

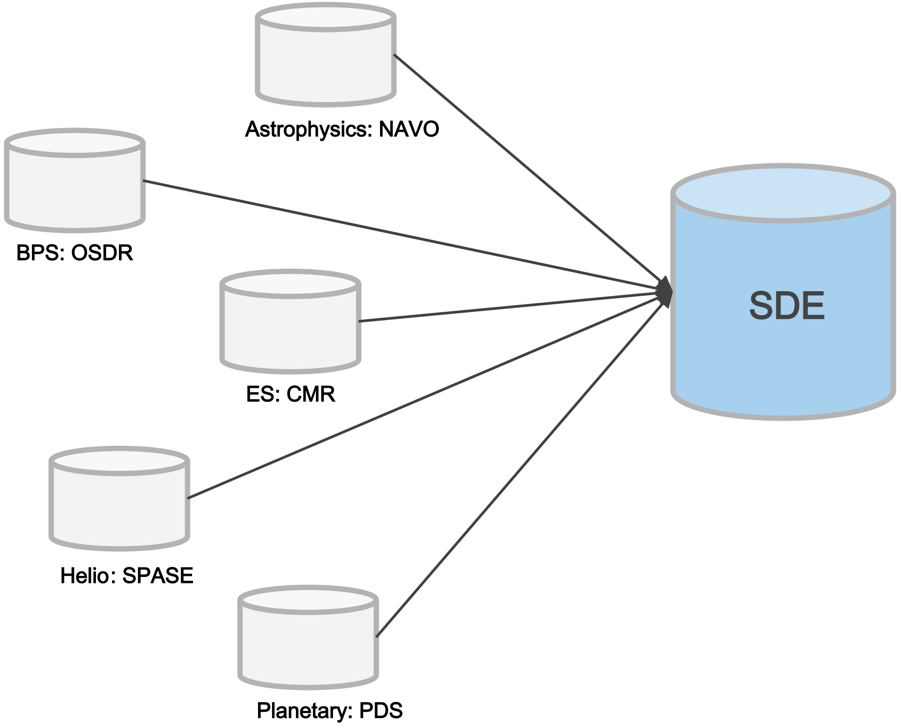
However, in order to do this, AI/ML models and datasets must be

- Described using a common yet extensible metadata model
- Cataloged into a centralized location(s)

So that the SDE and other capabilities can index the content.

Advancing Analyses for New Science Discoveries

Currently the SDE indexes dataset metadata from centralized repositories. Incorporation of AI/ML datasets and models into a centralized repository would streamline integration into capabilities like the SDE.



*Code & documentation comes from a variety of other sources.

Discussion

How can we develop a cohesive strategy for creating and managing AI/ML metadata across SMD?

How do we identify or curate AI/ML resources across SMD?

- There's a risk of information sprawl
- To enable discovery, developing a centralized repository of AI/ML models and datasets would be helpful
- This can be at the division level or SMD level– requires community discussion

Thank you!

Contact me: Kaylin.m.bugbee@nasa.gov