Transform Your Search Capability Using Insight Engines

February 28, 2024

Kaylin Bugbee, Applications Data Manager at NASA Marshall Space Flight Center

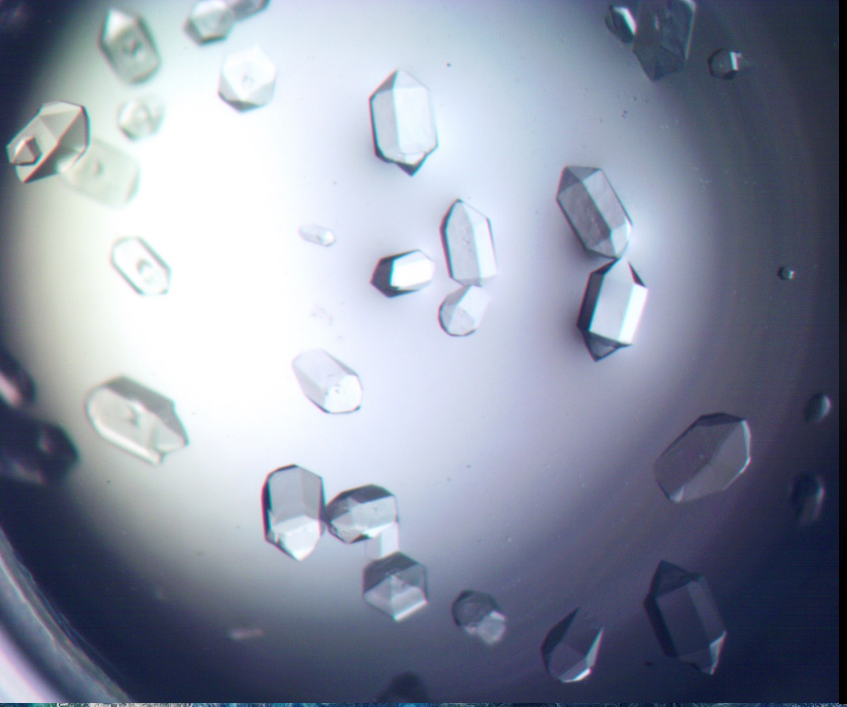
Kaylin Bugbee
NASA Marshall Space Flight Center

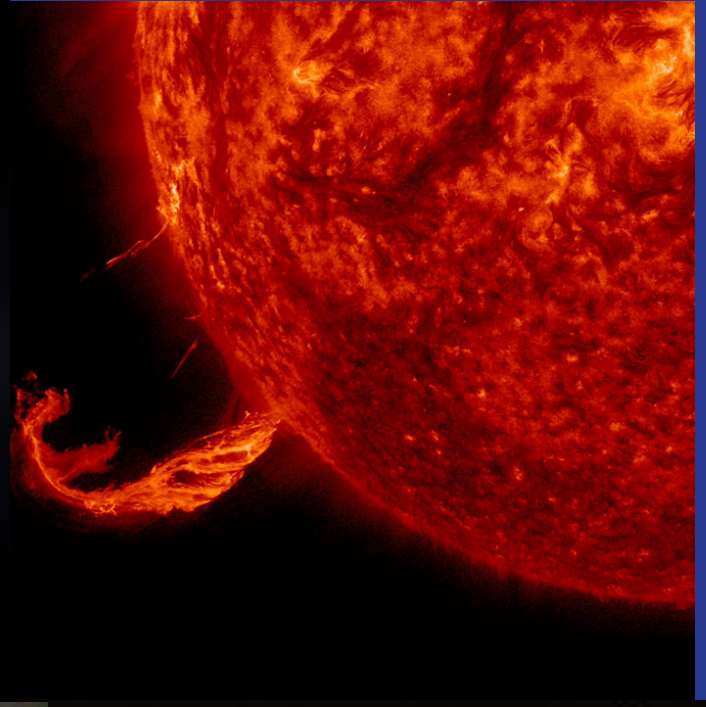
- Supported a number of key initiatives for NASA including:
  - The Climate Data Initiative, a key component in President Obama's Climate Action Plan
  - The Multi-Mission Algorithm and Analysis Platform, a collaborative project between NASA and the European Space Agency to enable science on the cloud
  - Metadata quality review project to review all 7000+ of NASA's Earth science metadata records and provide actionable recommendations for improvement
  - Transform to Open Science (TOPS) Project Scientist for the TOPS Project Office at MSFC
- Major publications include:
  - Bugbee, K., D. Smith, S. Wingo, and E. Foshee (2023), The art of scientific curation, Eos, 104, https://doi.org/10.1029/2023EO230201. Published on 19 May 2023.
  - Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From open data to open science. Earth and Space Science, 8, e2020EA001562. https://doi.org/10.1029/2020EA001562
- M.S. in Earth System Science from the University of Alabama in Huntsville

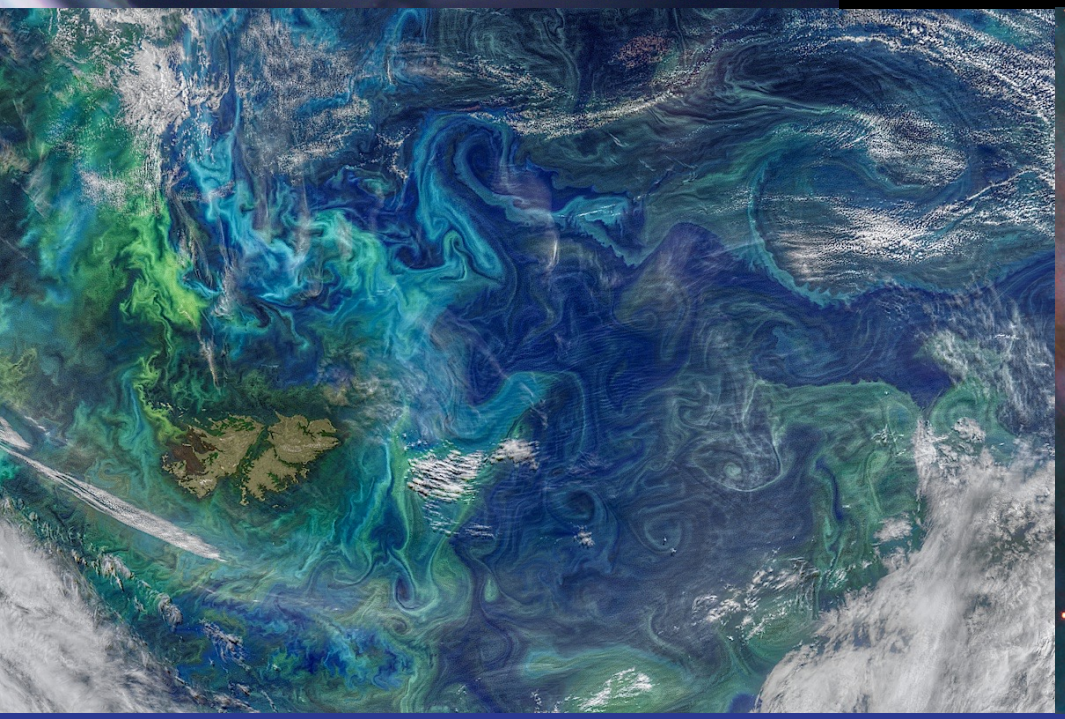
“The National Aeronautics and Space Administration (NASA) has become a **knowledge agency.** Long after the Mars Surveyor has gone silent, Hubble has met the same fate as Mir, and the Moderate Resolution Imaging Spectroradiometer has produced its final set of images, what will endure are the **volumes of valuable data** that these instruments and many others have collected over their lifetimes.”

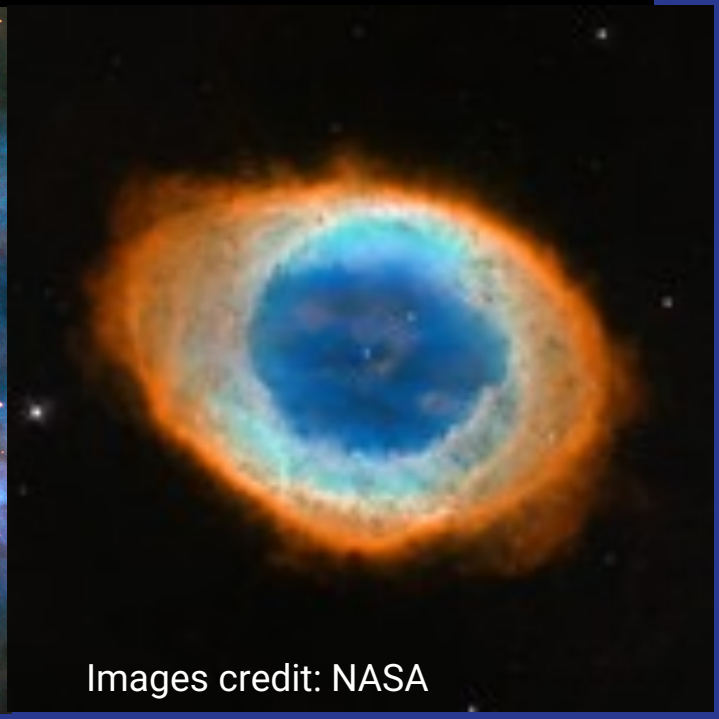
National Research Council. 2002. Assessment of the Usefulness and Availability of NASA's Earth and Space Science Mission Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/10363.

Image credit: NASA



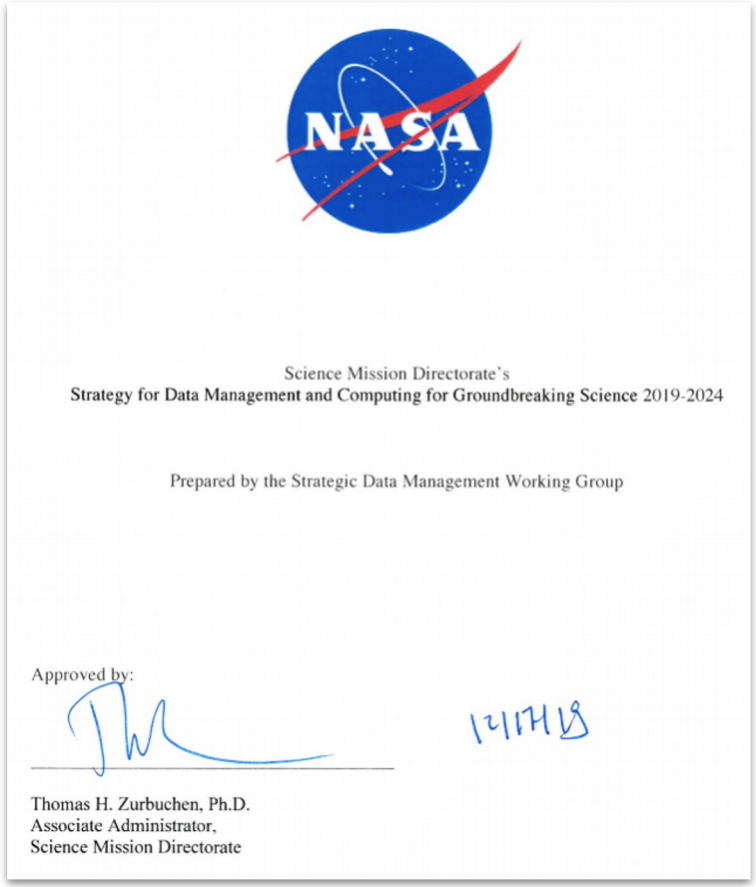






Images credit: NASA

Open Science at NASA



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:

12/17/19

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

Goal 1: Develop and Implement Capabilities to Enable Open Science

1.3: Develop and implement an SMD data catalog to support **discovery** and **access** to complex scientific data **across divisions**

- NASA SMD Strategy for Data Management and Computing for Groundbreaking Science
- Open Source Science Initiative (OSSI) - "commitment to building an inclusive, open science community"

Science Discovery Engine (SDE)

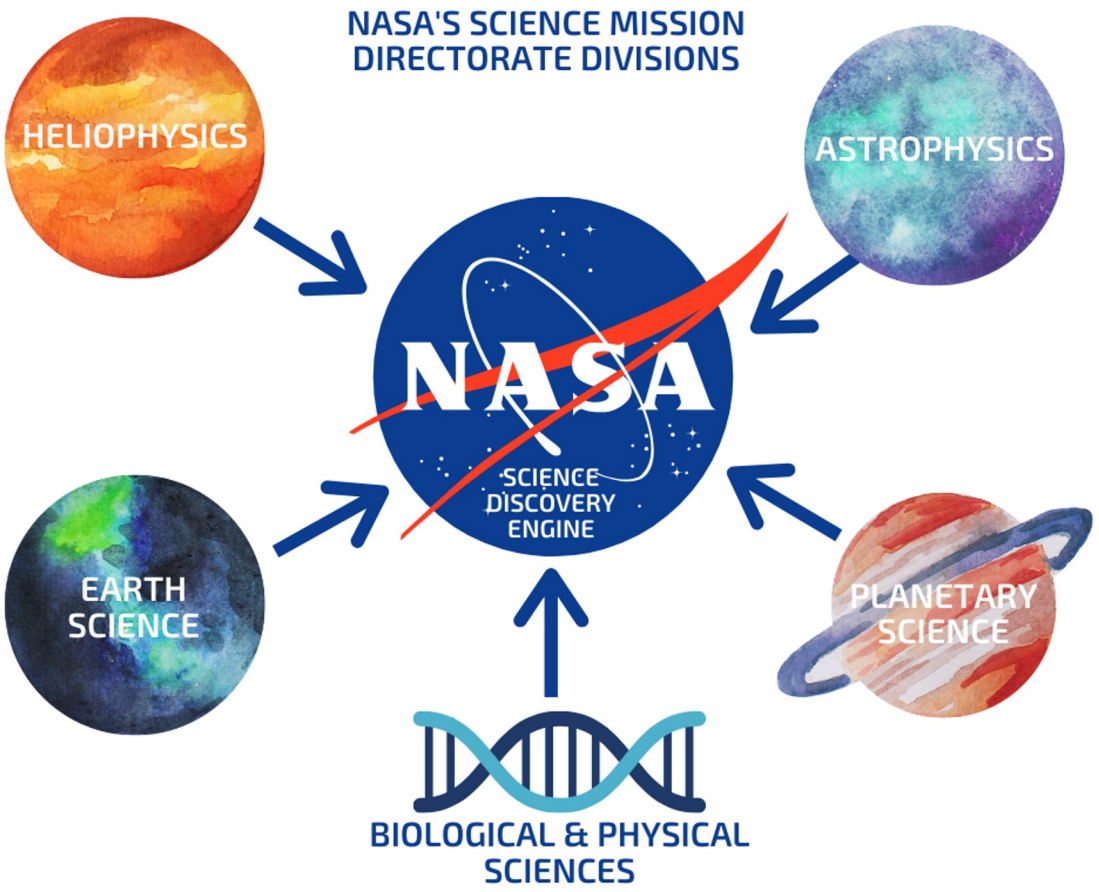


Image Credit: SDE team

Goals:

1. Enable rapid discovery of NASA's open science data, software and documentation
2. Support NASA's open science goals and infrastructure
3. Promote interdisciplinary science
4. Prototype emerging technologies and search techniques including Large Language Models

Access the SDE at:
https://sciencediscoveryengine.nasa.gov/
Or use the QR code.

SDE Scope

Includes metadata about science data but also key contextual information such as code, imagery, tutorials, mission information.

The ability to search for all of NASA's open science data and supporting contextual information in a unified interface is unprecedented.

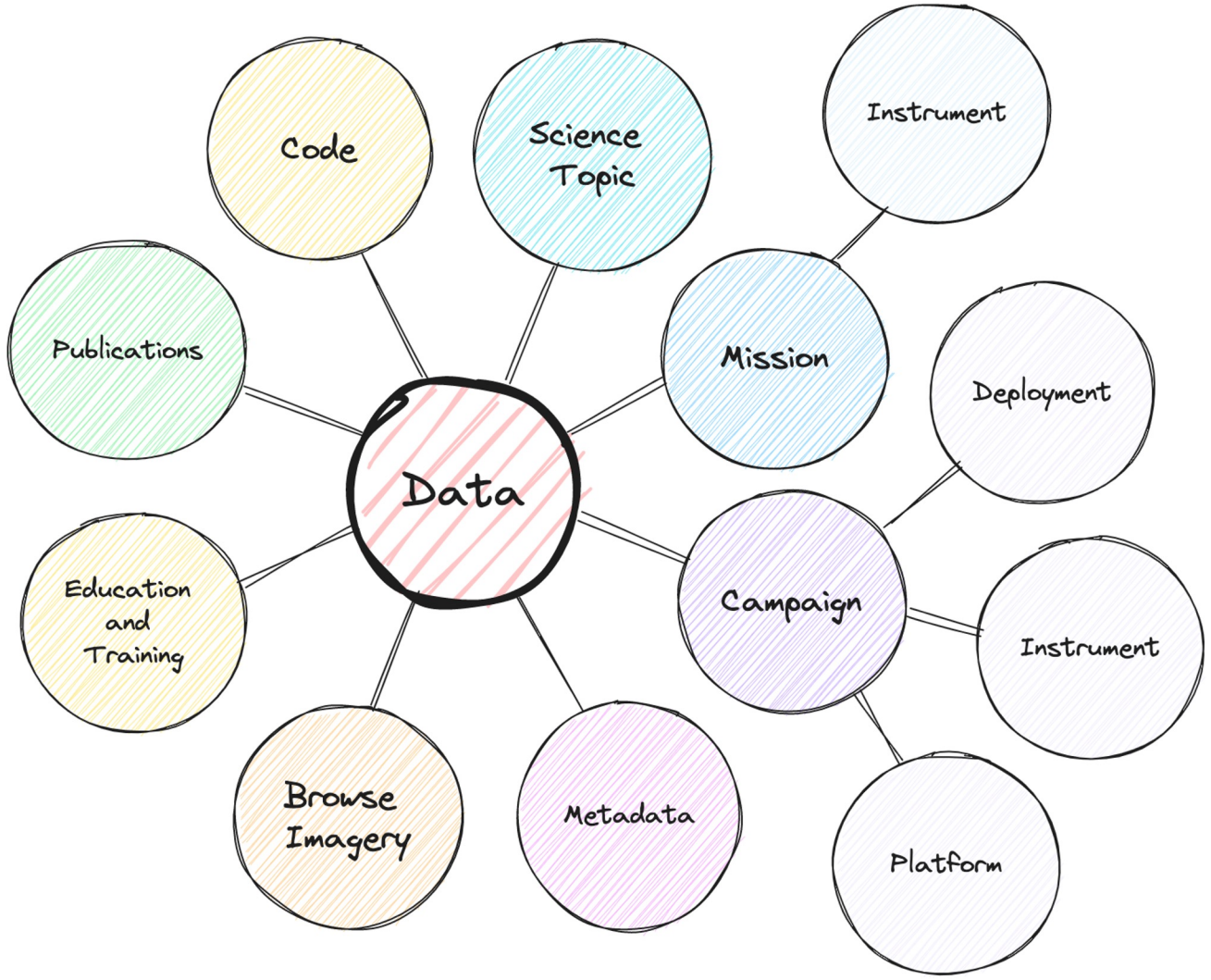


Image Credit: SDE team

SDE Technical Implementation

The SDE uses a technology called an **Insight Engine** to bring heterogeneous content together.

- ✓ Insight Engines apply relevancy methods to describe, discover, organize, and analyze data
- ✓ Can incorporate structured information (metadata about data), and unstructured information (web pages, algorithm documentation, user guides, code and tutorials)
- ✓ Enables richer indexes, more complex queries, elaborated relevancy methods, and multiple touchpoints to support users
- ✓ Allows for specialized user interface customization to meet the needs of the scientific community
- ✓ Leverages NLP techniques
- ✓ Using an Insight Engine technology called Sinequa for the SDE

Current Functional Design

- Leverages **existing data and information** to increase discovery, access and reuse across divisions
- Curation management leverages custom built web app for **curation** of HTML sources
- Uses Sinequa Insight Engine capability for search
  - Provides an **API** that can be used for direct querying and powers the front end
- Front end **tailored for science**
  - Science facets such as platforms, instruments
  - Filter by research object type (data, images, documents)
  - Knowledge panels and 360 views to explore science information (in development now)

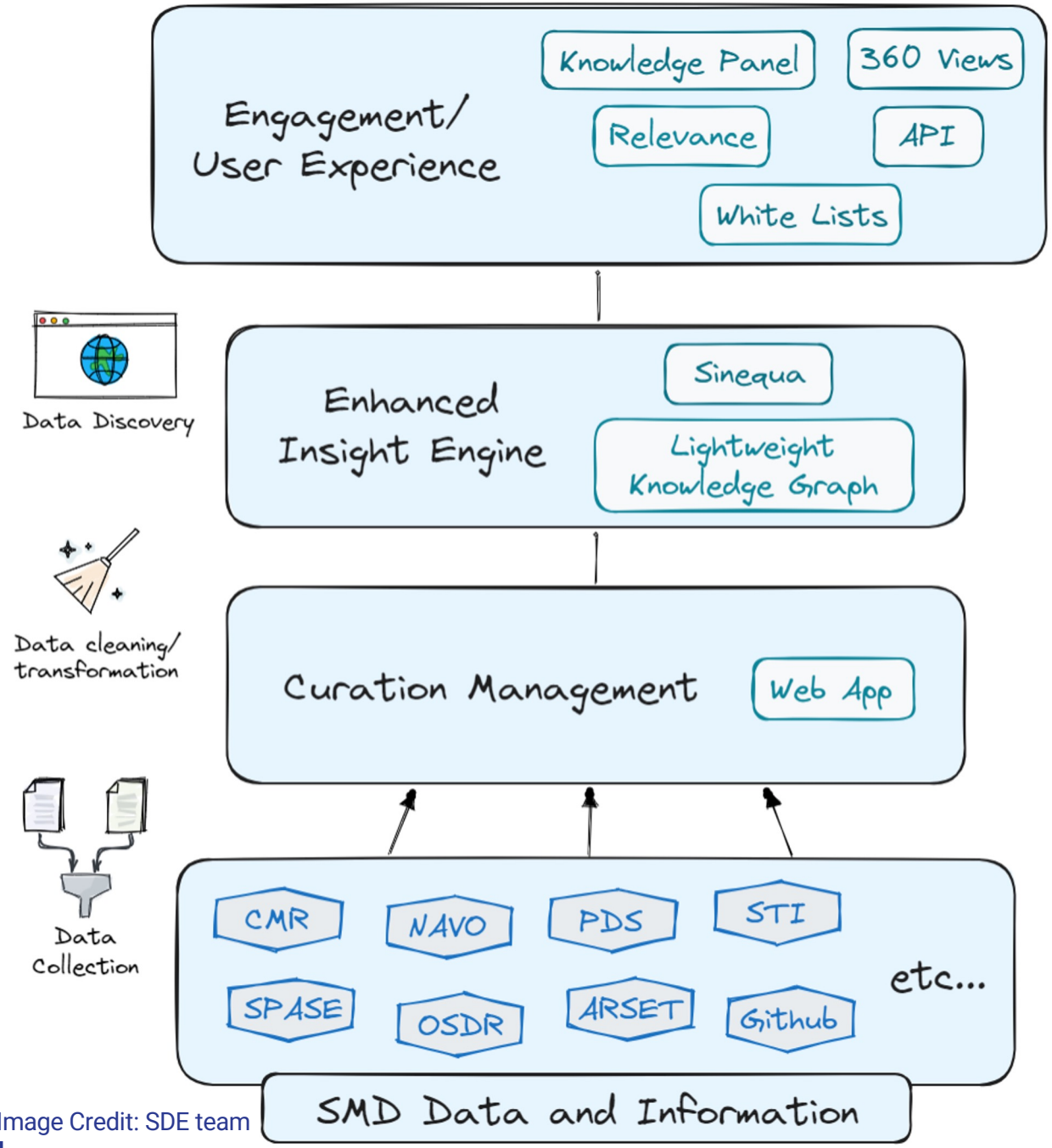


Image Credit: SDE team

SDE Metrics To Date

Numbers as of January 2024

Metrics are constantly evolving as curation workflows improve and new information is added.

Sources are refreshed or reindexed at the following frequencies:

- Daily for dataset metadata
- Monthly for all other content (websites)

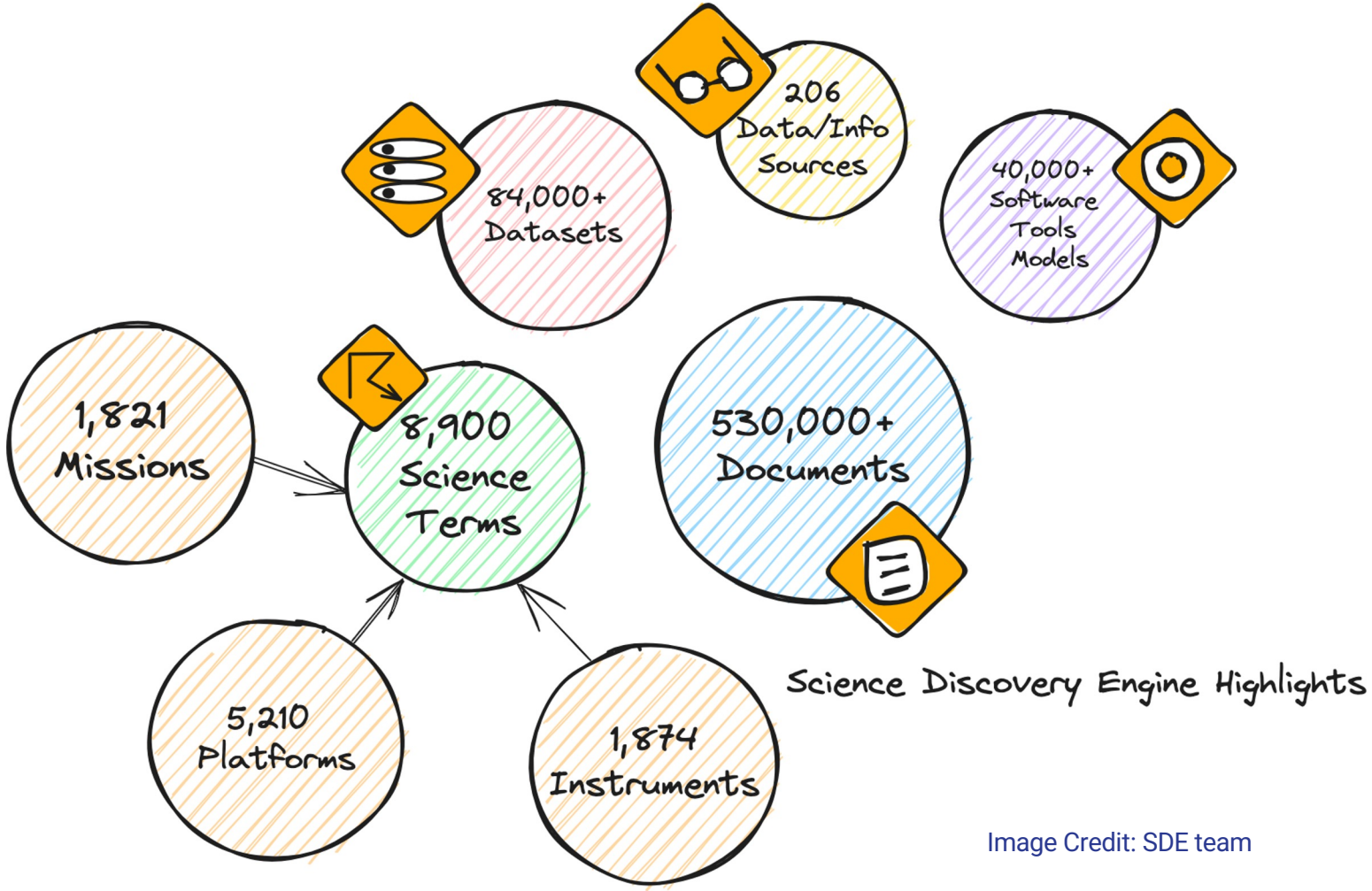
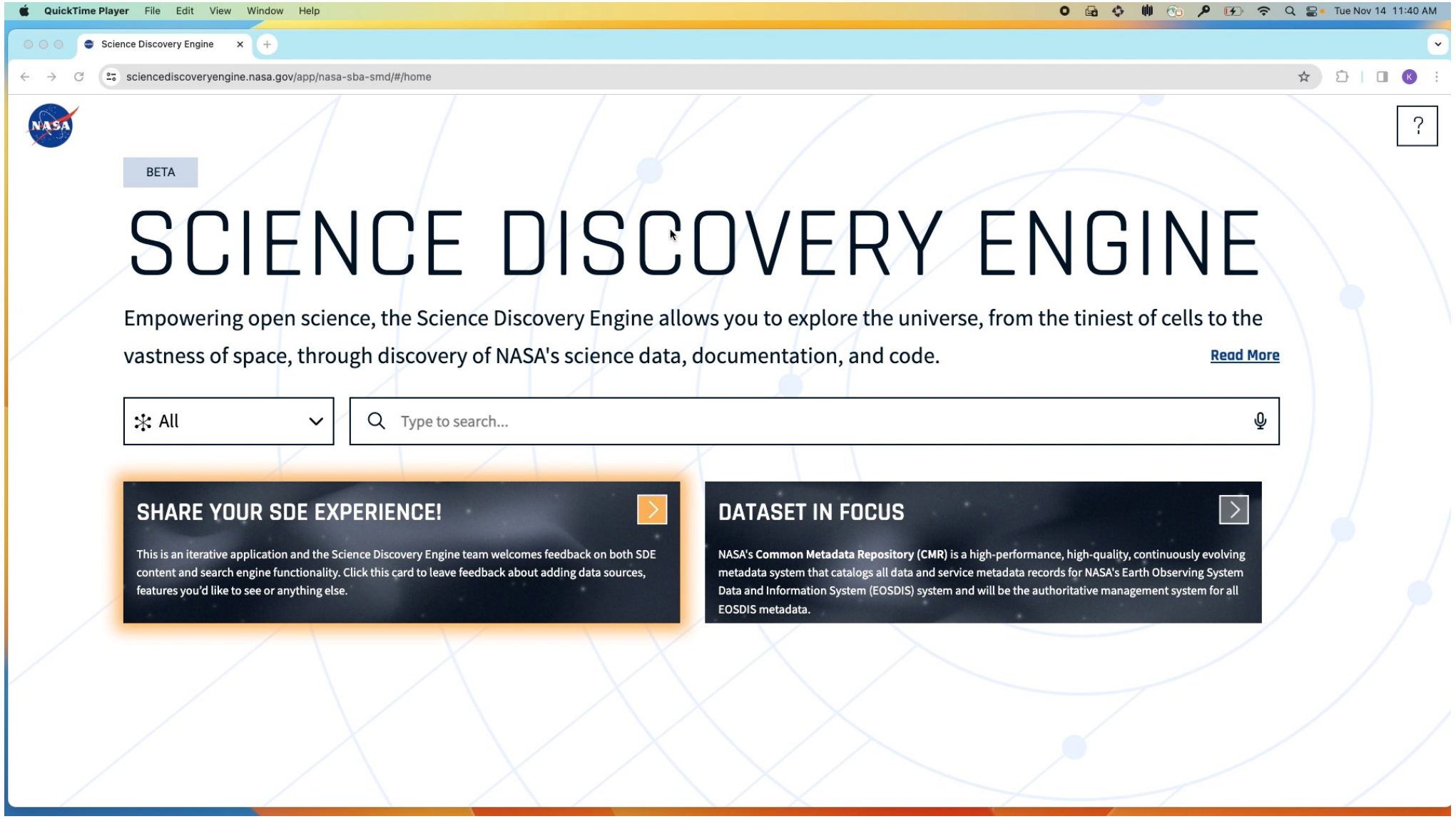


Image Credit: SDE team

SDE: Demo


QuickTime Player File Edit View Window Help
Tue Nov 14 11:40 AM
Science Discovery Engine
sciencediscoveryengine.nasa.gov/app/nasa-sba-smd/#/home
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.
Read More
All
Type to search...
SHARE YOUR SDE EXPERIENCE!
This is an iterative application and the Science Discovery Engine team welcomes feedback on both SDE content and search engine functionality. Click this card to leave feedback about adding data sources, features you'd like to see or anything else.
DATASET IN FOCUS
NASA's Common Metadata Repository (CMR) is a high-performance, high-quality, continuously evolving metadata system that catalogs all data and service metadata records for NASA's Earth Observing System Data and Information System (EOSDIS) system and will be the authoritative management system for all EOSDIS metadata.

Upcoming Planned Features and Enhancements

- Continue incorporating new NASA science content into the SDE
- Provide access to an open API and supporting documentation
- Integrate with NASA’s science literature portal
- Enable 360 views into key aspects of the data
- Allow customers to build specialized scientific search applications using the content and infrastructure made available via the SDE

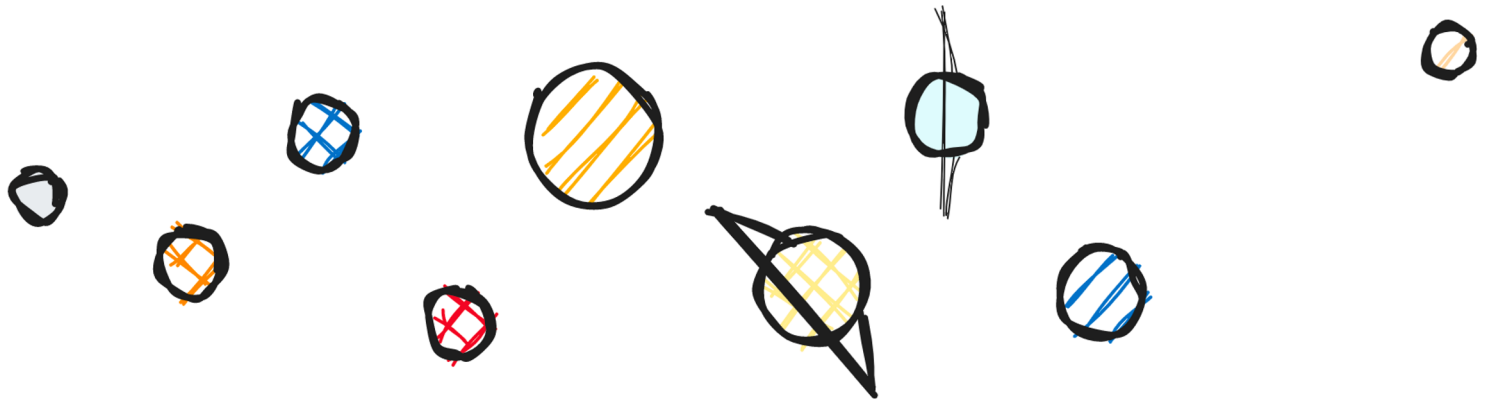


Image Credit: SDE team

AI Use Cases

The search paradigm is changing – generative AI and large language models (LLMs) are changing how people search for data and information. While there will always be a need for keyword search, users also want to have a conversation or dialog when discovering and understanding data and information.

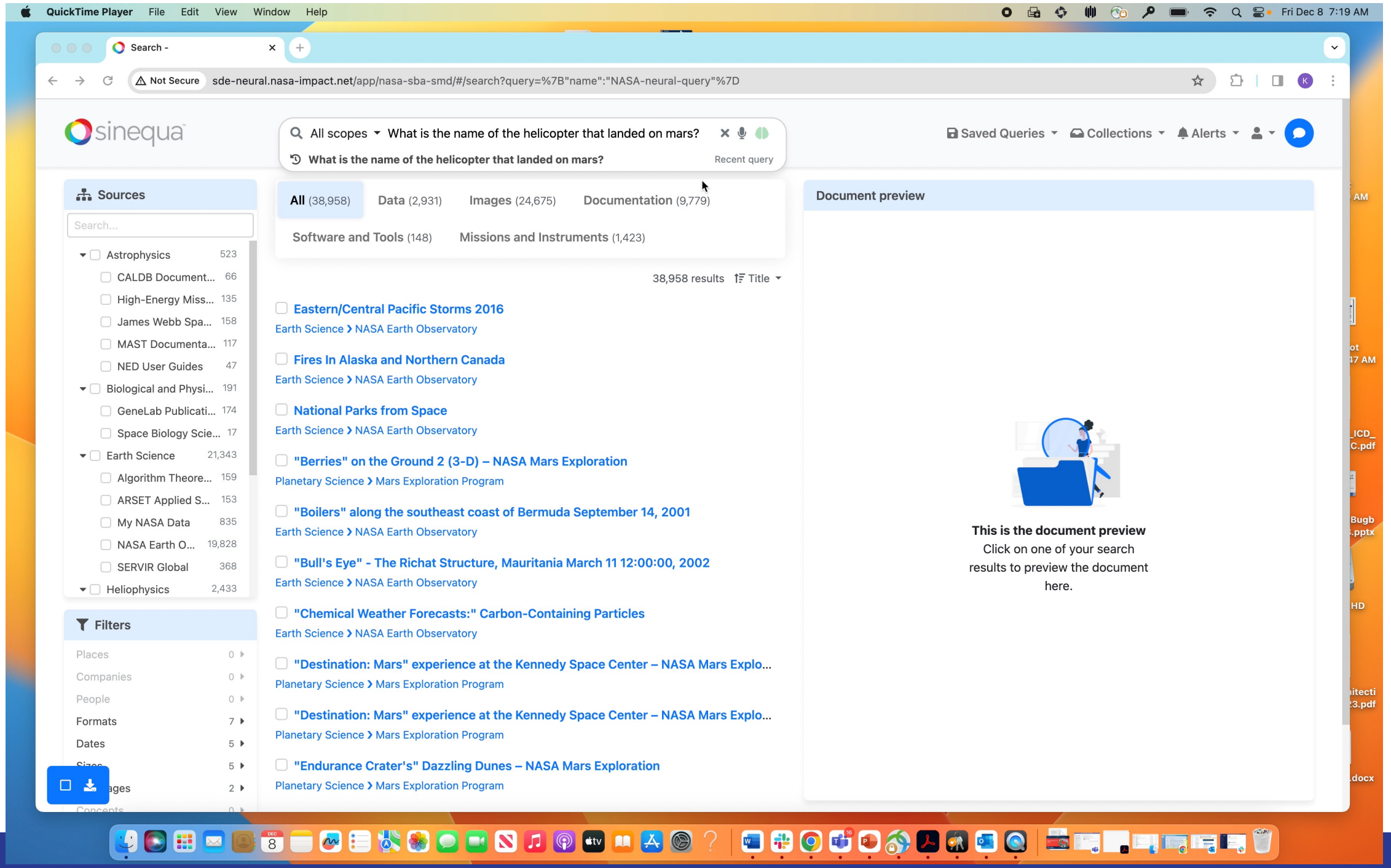
The SDE team is prototyping a ‘conversational’ search capability using:

- Technology from Sinequa called Neural Search
- Encoder only Large Language Model developed for SMD

It is important to test the reliability, trustworthiness and accuracy of these types of capabilities in addition to understanding cost and architecture considerations.

We are also looking at how we can leverage LLMs to streamline curation and data workflows by:

- Creating structured metadata from unstructured content
- Classifying documents
- Enabling an LLM based data layer that brokers API calls for users


QuickTime Player File Edit View Window Help
Fri Dec 8 7:19 AM
Search -
Not Secure
sde-neural.nasa-impact.net/app/nasa-sba-smd/#/search?query=%7B"name":"NASA-neural-query"%7D
sinequa
All scopes
What is the name of the helicopter that landed on mars?
What is the name of the helicopter that landed on mars?
Recent query
Saved Queries
Collections
Alerts
Sources
Search...
Astrophysics 523
CALDB Document... 66
High-Energy Miss... 135
James Webb Spa... 158
MAST Documenta... 117
NED User Guides 47
Biological and Physi... 191
GeneLab Publicati... 174
Space Biology Scie... 17
Earth Science 21,343
Algorithm Theore... 159
ARSET Applied S... 153
My NASA Data 835
NASA Earth O... 19,828
SERVIR Global 368
Heliophysics 2,433
Filters
Places 0
Companies 0
People 0
Formats 7
Dates 5
5
2
All (38,958)
Data (2,931)
Images (24,675)
Documentation (9,779)
Software and Tools (148)
Missions and Instruments (1,423)
38,958 results
Title
Eastern/Central Pacific Storms 2016
Earth Science › NASA Earth Observatory
Fires In Alaska and Northern Canada
Earth Science › NASA Earth Observatory
National Parks from Space
Earth Science › NASA Earth Observatory
"Berries" on the Ground 2 (3-D) – NASA Mars Exploration
Planetary Science › Mars Exploration Program
"Boilers" along the southeast coast of Bermuda September 14, 2001
Earth Science › NASA Earth Observatory
"Bull's Eye" - The Richat Structure, Mauritania March 11 12:00:00, 2002
Earth Science › NASA Earth Observatory
"Chemical Weather Forecasts:" Carbon-Containing Particles
Earth Science › NASA Earth Observatory
"Destination: Mars" experience at the Kennedy Space Center – NASA Mars Explo...
Planetary Science › Mars Exploration Program
"Destination: Mars" experience at the Kennedy Space Center – NASA Mars Explo...
Planetary Science › Mars Exploration Program
"Endurance Crater's" Dazzling Dunes – NASA Mars Exploration
Planetary Science › Mars Exploration Program
Document preview
This is the document preview
Click on one of your search results to preview the document here.

Benefits to This Approach

- A unified search capability makes NASA's open science data and information more visible, accessible, usable and valuable
- An insight engine type capability can be implemented with minimal impact to organizational data stewards and data providers
    - Content can be indexed without the need to complete time and labor-intensive activities like schema matching, etc.…
    - Did not require data stewards and providers to do additional work to have their content included in the search capability
- An insight engine can be customized as needed for specific use cases, providing value to stakeholders who are searching for data and information

What Did It Take to Make This Happen?

- A cross-disciplinary team of scientists, informatics specialists, developers and technology specialists to build and maintain the SDE.
- A Sinequa senior solutions engineer embedded in the team to help provide detailed Sinequa guidance whenever the team had issues.
- An interdisciplinary advisory board (or working group) with members from each of the NASA science topic areas to provide input and guidance on content to be included and features to be built.
- Support from the agency's OCIO to acquire our part of the Sinequa license and to help us migrate the architecture to the cloud.
- Support from key leadership at NASA headquarters.
- A year to build and populate the beta SDE which was released in December 2022.

Lessons Learned

- **Using an insight engine for a domain-specific use case required custom development**. This included:
  - Creating a customized user interface to meet the needs of the scientific community
  - Identifying relevant facets and building the corresponding term lists needed to create those facets
  - Developing documentation and tools to make the API more friendly for the science community
  - Creating an application to help in managing curated indexing workflows
- **Content curation is essential for a focused, domain-specific use case.**
  - Blindly indexing everything degraded the search experience for scientists
  - Curation
    - Reduces the likelihood of duplicative content
    - Removes content of no value to the user community (HTML styling information, etc)
    - Removes content that is not searchable in a text-based search environment
    - Reduces the number of documents to be managed

Thank you!

Kaylin.m.bugbee@nasa.gov