

The Science Discovery Engine: Connecting Heterogeneous Scientific Data and Information

Kaylin Bugbee¹, Ashish Acharya², Emily Foshee²,
Muthukumaran Ramasubramanian², Carson Davis²,
Bishwas Praveen², Kartik Nagaraja³, Shravan
Vishwanathan³, Stephanie Wingo², and Rachel
Wyatt²

1: NASA Marshall Space Flight Center

2: University of Alabama in Huntsville

3: Left Right Mind



Science Discovery Engine (SDE)

- The SDE is a source of trusted and curated open science data and information
- The SDE includes:
 - Metadata about science data
 - Code
 - Documentation
 - Images
 - Tutorials
 - Mission and instrument information
- Includes over 84,000 metadata records and thousands of documents from hundreds of sources across NASA science



Access at:

<https://sciencediscoveryengine.nasa.gov/>

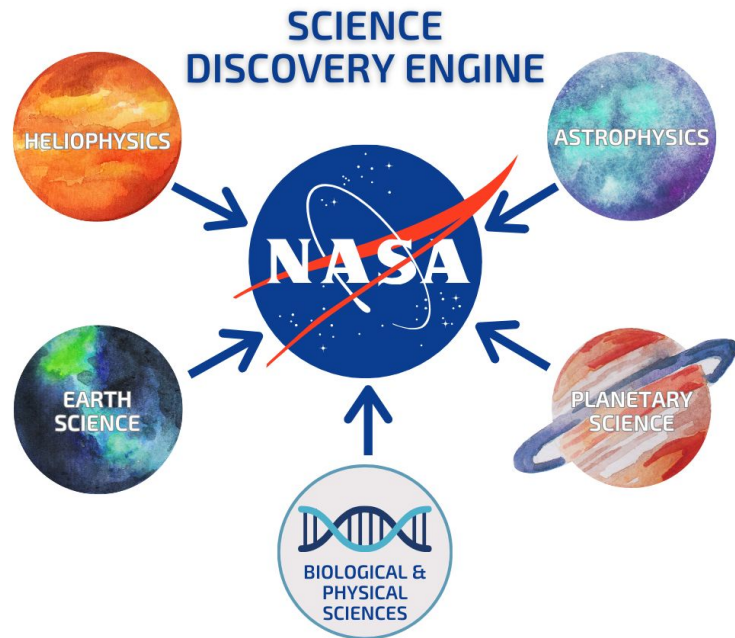
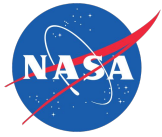
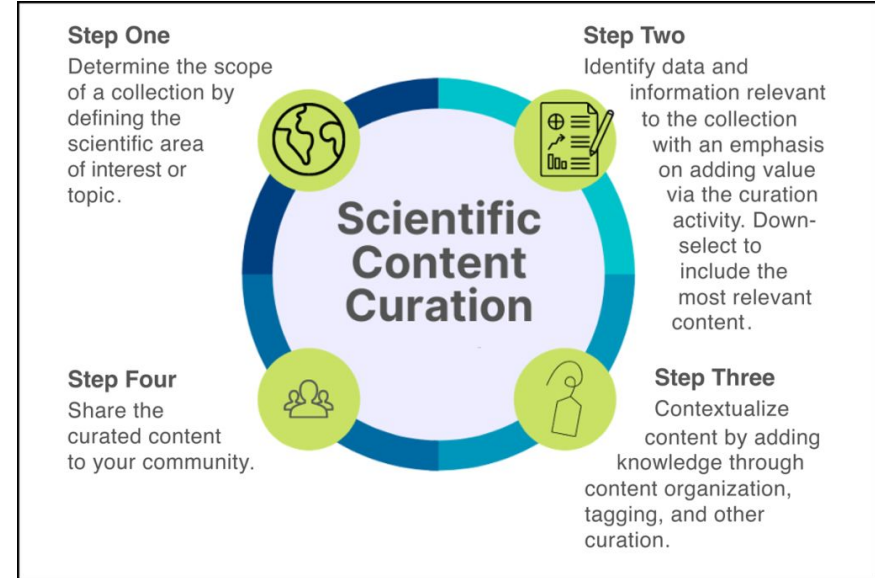


Image credit: SDE team

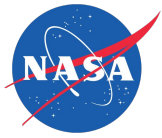


What is Scientific Content Curation?

- Act of involving authoritative experts who identify, gather, validate, synthesize, organize, and present contextual details necessary to discover, understand, and use scientific data and knowledge effectively
- A value-adding effort that includes any activity or process that accelerates progress toward actionable science or makes it easier for diverse audiences to digest scientific information

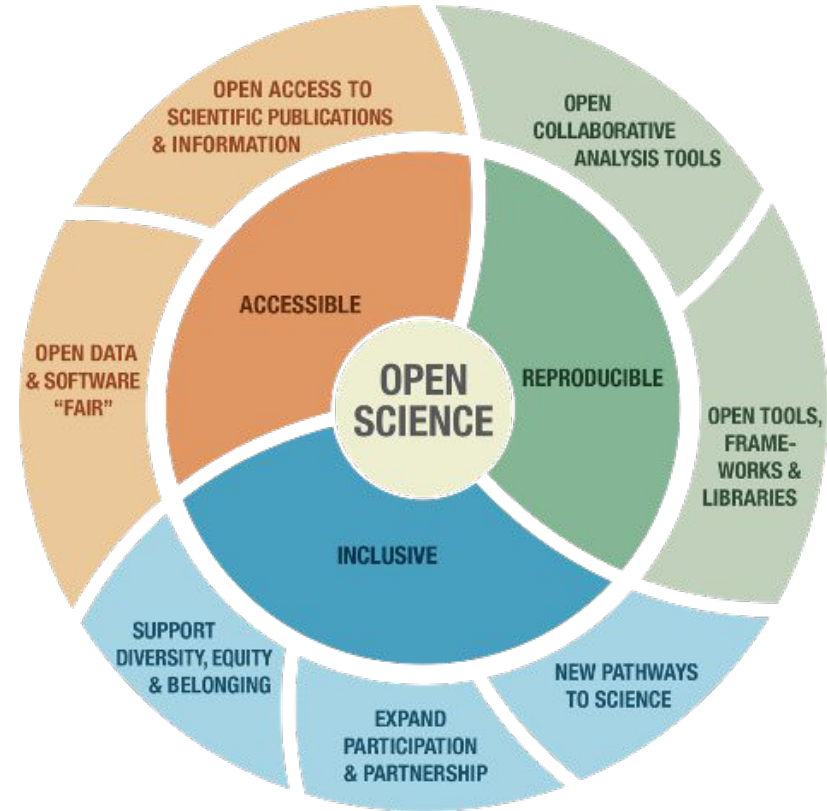


Citation: Bugbee, K., D. Smith, S. Wingo, and E. Foshee (2023), The art of scientific curation, Eos, 104, <https://doi.org/10.1029/2023EO230201>. Published on 19 May 2023.



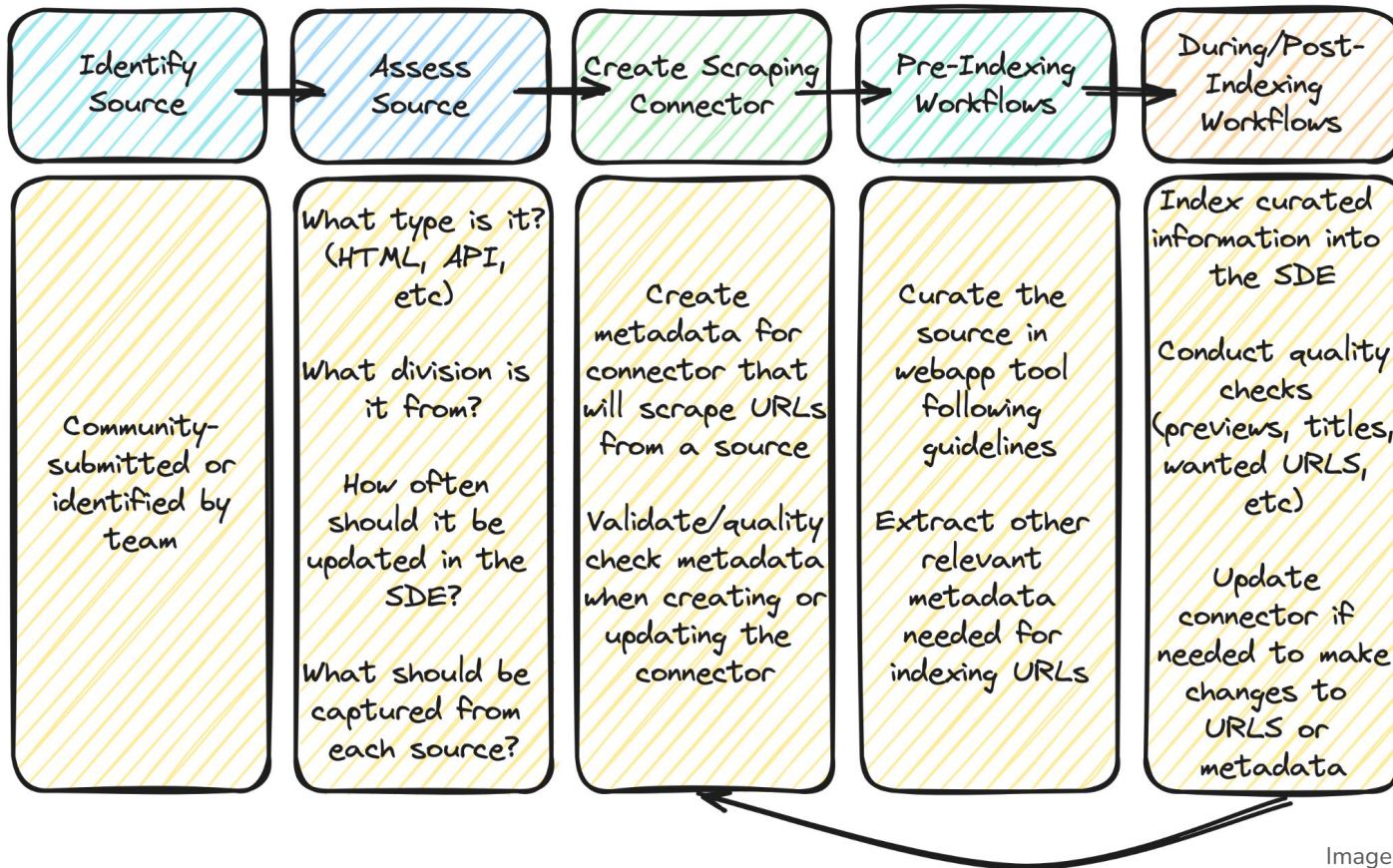
Why Do We Need Scientific Content Curation?

- Open Data, Software, Access
 - More data, code, documentation and publications will be shared or made openly available as more people embrace open science principles
 - We need a way to help everyone find the information they are looking for
- As science becomes more inclusive, we need to provide more equitable access to data and information
 - Curated content for different audiences and scientific communities will be needed to help support belonging and to expand participation
- Curated information can be used to build trusted generative AI applications





Scientific Content Curation in the SDE





Scientific Content Curation in the SDE

- When curating individual URLs from a specified source, need to decide:
 - Is it scientifically relevant?
 - Does it have searchable text?
 - Is the current title unique and descriptive?
- Typically during the curation process we exclude:
 - Explicit data files
 - Code without context
 - XML or other metadata-type files
 - Image-only URLs
 - Compressed files
 - Links that redirect or are broken
- When assigning a document type, need to decide:
 - What is the category that describes the URL as a whole?
 - One URL may cover multiple document types

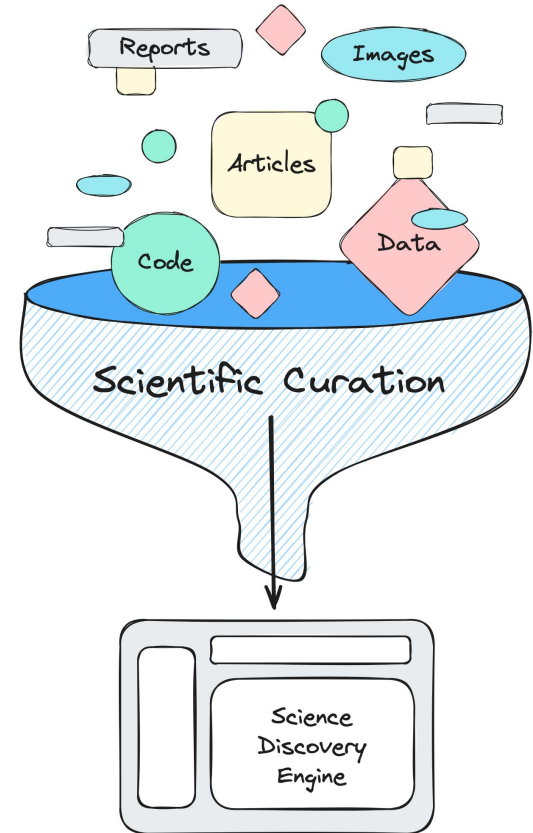


Image credit: SDE team



Scientific Content Curation in the SDE

- Metadata schemas are NOT consistent across each of the five SMD divisions
- In order to display metadata in a simplified and cohesive manner, it was required to:
 - Find common metadata elements between all divisions
 - Create our own SDE 'metadata schema' based on the common elements
- Using the SDE metadata schema:
 - Generated dataset landing pages for datasets within each of the divisional metadata aggregators (SPASE, CMR, NAVO, OSDR, PDS)

Yohkoh HXT Solar Images



Overview

Details

Links

Scientific Focus

Heliophysics

Agency

JAXA

Repository

MSU

Persistent Identifier

-

Description

Yohkoh HXT (Hard X ray Telescope) Solar Images. HXT had the following characteristics: effective area: 70 sq. cm, time resolution of 2 s, angular resolution ~5 arc s FWHM, energy range 15-100 keV

Data Format

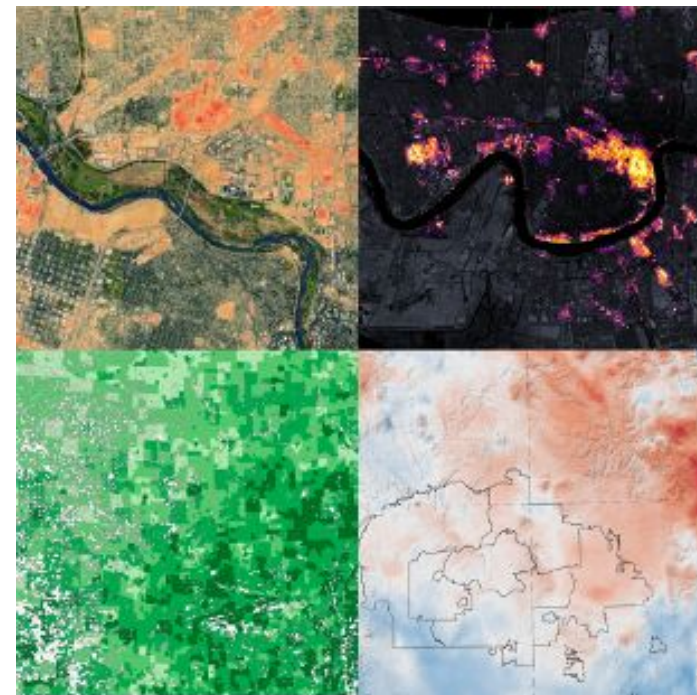
Access Constraints

Image credit: SDE team



SDE: Enabling Curated Search Applications

- Scientific communities have a need to curate and share data and information around their specific use case.
- Prototype NASA Environmental Justice Data Search portal supports EJ communities by expanding awareness, accessibility, and use of Earth science data by providing a curated, trusted, easy-to-use search interface.
- Curated data catalog serves as the basis for the initial collection
- Data and information organized by:
 - Environmental Justice indicator (such as disasters, urban flooding, extreme heat, and food availability)
 - Possible intended use cases for each dataset such as 'explore data' or 'analyze data'



Examples of how NASA Earth science data are being applied to environmental justice. Clockwise from upper left: Identifying vulnerable populations using remotely-sensed urban heat island data in Sacramento, CA; detecting power outages using nighttime lights imagery in New Orleans, LA; assessing Navajo Nation water resources using NASA's Drought Severity Evaluation Tool (DSET) in the U.S. Southwest; helping policy officials identify communities in need using U.S. Social Vulnerability Index data from NASA's Socioeconomic Data and Applications Center (SEDAC). Credit: NASA ESDS.



SDE

?

BETA

ENVIRONMENTAL JUSTICE DATA SEARCH INTERFACE

🔍 Type to search...



Search

In order to search within a focus area, type your query and click the focus area below.
Or, click a focus area to view all datasets within.

CLIMATE CHANGE

DISASTERS

EXTREME HEAT

FOOD AVAILABILITY

HUMAN DIMENSIONS

HEALTH & AIR QUALITY

URBAN FLOODING

WATER AVAILABILITY



Image credit: SDE team



Future of Content Curation in the SDE

- The scientific curation process within the SDE requires substantial human involvement. We want to automate processes wherever possible.
- Potential automated workflows include:
 - Document-type tagging
 - Classification of documents for specialized search applications
 - Classification of documents for division type for sources that cover 2+ divisions
 - Grabbing titles from PDFs and other file types that have 'bad' scraped titles

From the Smallest Cells...

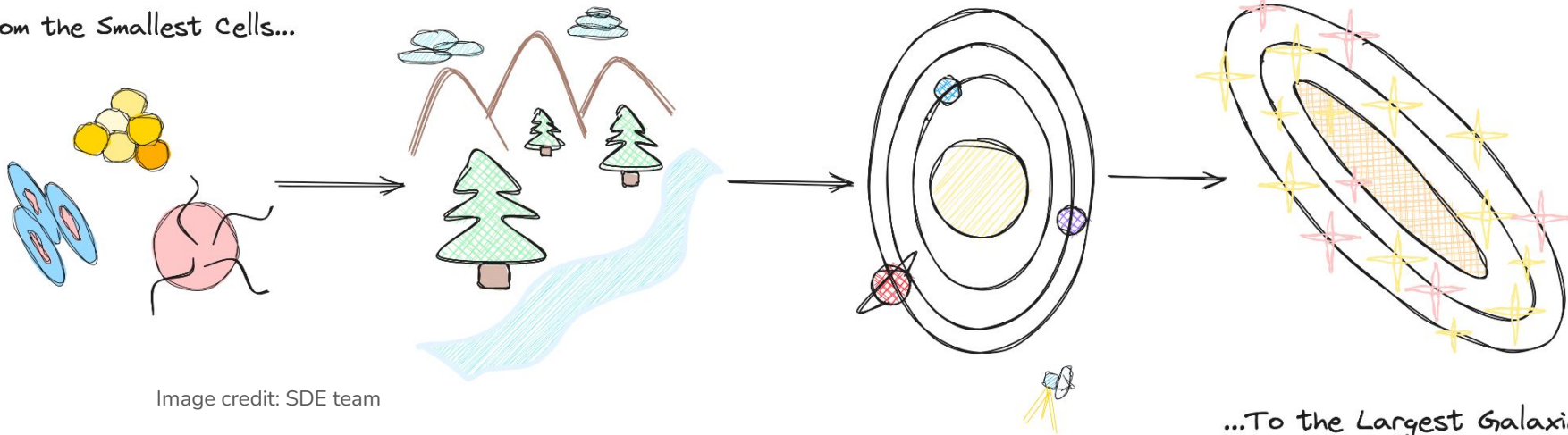


Image credit: SDE team

...To the Largest Galaxies

Thank you!

Science Discovery
Engine

