



Data Search and Exploration using EarthCube Data Discovery Studio



Data Search and Exploration using EarthCube Data Discovery Studio

Ilya Zaslavsky (1), David Valentine (1), Stephen Richard (2), Ouida Meier (3)

1: UC San Diego; 2: US Geoscience Information Network; 3: University of Hawaii



DDStudio: Beyond Search

EarthCube Data Discovery Studio is a platform for finding and exploring geoscience data.

DDStudio supports various modes of data discovery.

OPEN

Metadata Enhancement

DDStudio implements a metadata augmentation pipeline that uses text analytics and geoscience ontologies to generate keywords, spatial and temporal extents, and organization identifiers.

OPEN

Schema.org Markup

All metadata records in DDStudio are presented using schema.org markup and referenced in sitemaps submitted to Google. By now, Google indexed 900K records from DDStudio.

OPEN

Jupyter Integration

DDStudio prototyped workflows for a seamless transition from geoscience data discovery to research. Using the "Studio" link from search results in the discovery interface, users can launch Jupyter notebooks residing on

OPEN

Collection Management

DDStudio implemented a custom version of ESRi Geoportal Server, with the added capability that lets users save any discovered metadata records into collections. The found records can be added to new or existing collection. Users can also export collections and share them with collaborators, or import collections developed by others.

OPEN

Watch DataDiscoveryStudio.org in Action!

OPEN

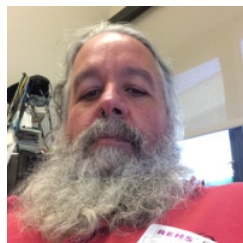
SGCI Collaboration

The Science Gateways Community Institute (SGCI) has served as an important partner to DDStudio over the last year.

OPEN

Ilya Zaslavsky (1), David Valentine (1), Stephen Richard (2), Ouida Meier (3)

1: UC San Diego; 2: US Geoscience Information Network; 3: University of Hawaii



PRESENTED AT:

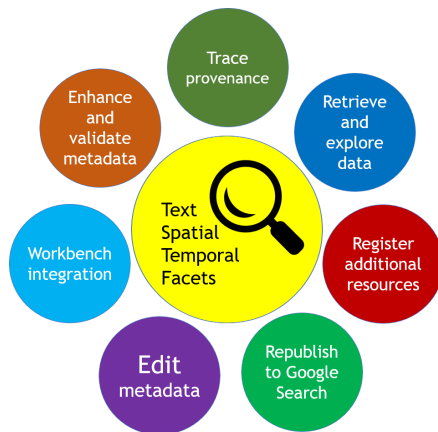


2020 EarthCube Annual Meeting

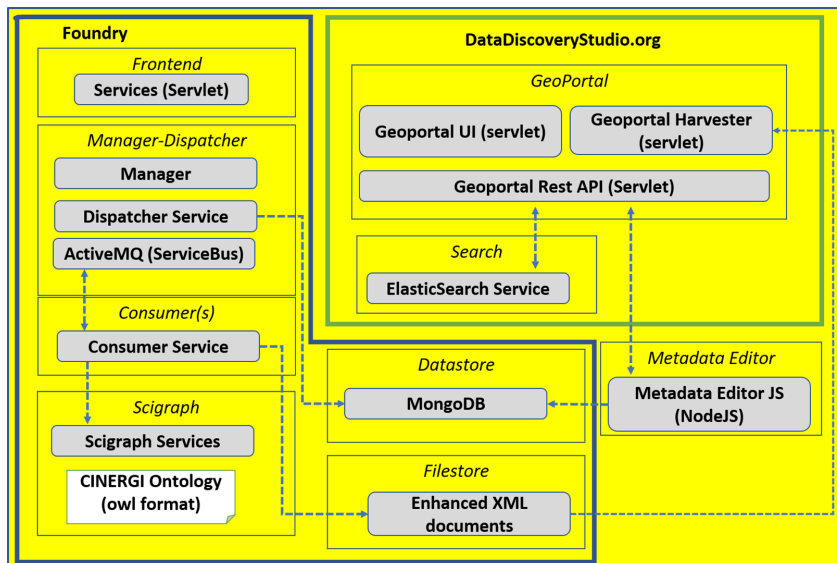
Virtual – June 18, 2020

DDSTUDIO: BEYOND SEARCH

EarthCube Data Discovery Studio is a platform for finding and exploring geoscience data.



Key DDStudio components

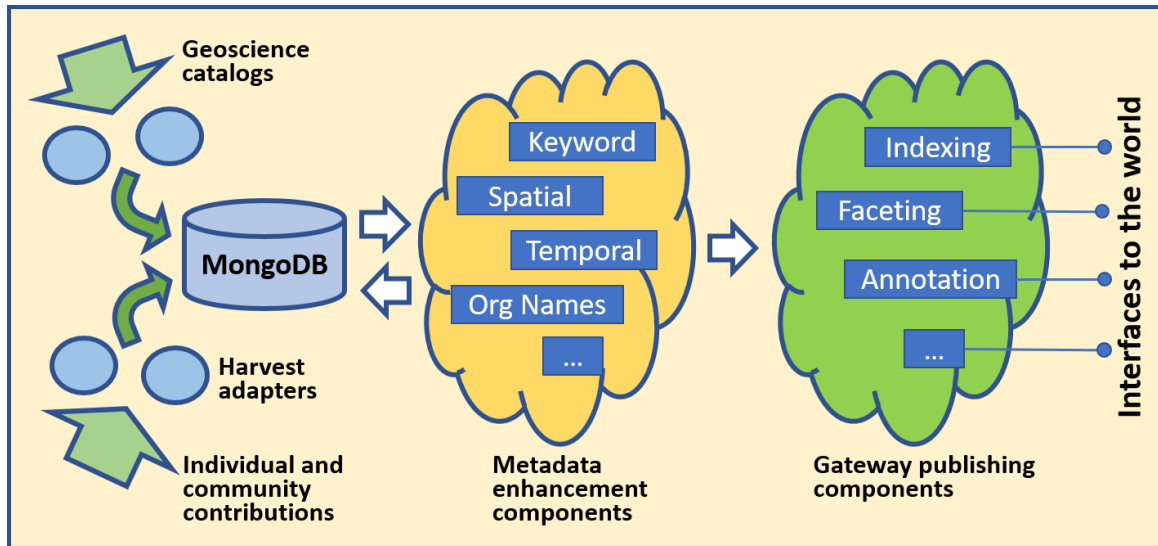


DDStudio is deployed on a set of virtual machines in OpenStack cloud platform. The key component is the DDStudio Foundry, which implements the metadata enhancement pipeline using service messaging bus, a producer/consumer architecture, and a MongoDB data store. Additional virtual machines are used for the user portal, Elasticsearch, and the Metadata editor.

Acknowledgment. NSF support (awards 1639764, 1639775) is gratefully acknowledged.

METADATA ENHANCEMENT

DDStudio implements a metadata augmentation pipeline that uses text analytics and geoscience ontologies to generate keywords, spatial and temporal extents, and organization identifiers.

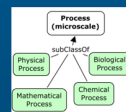
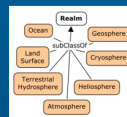


The keyword enhancer relies on GeoSciGraph services operating over 20+ geoscience ontologies, reconciled and organized into a faceted search hierarchy.

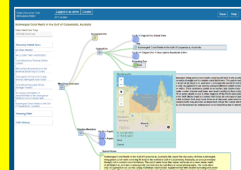
GeoSciGraph ontology management system provides semantic infrastructure. It relies on a cross-domain ontology of geoscience terms, integrating several independently developed ontologies or taxonomies

Some included ontologies:

- SWEET
- ENVO
- CHEBI
- YAGO (geo features)
- NASA GCMD (equipment, providers)
- GeoSciML
- Geochronology
- EDAM Bioinformatics (software terms and operations)
- Also: VIAF



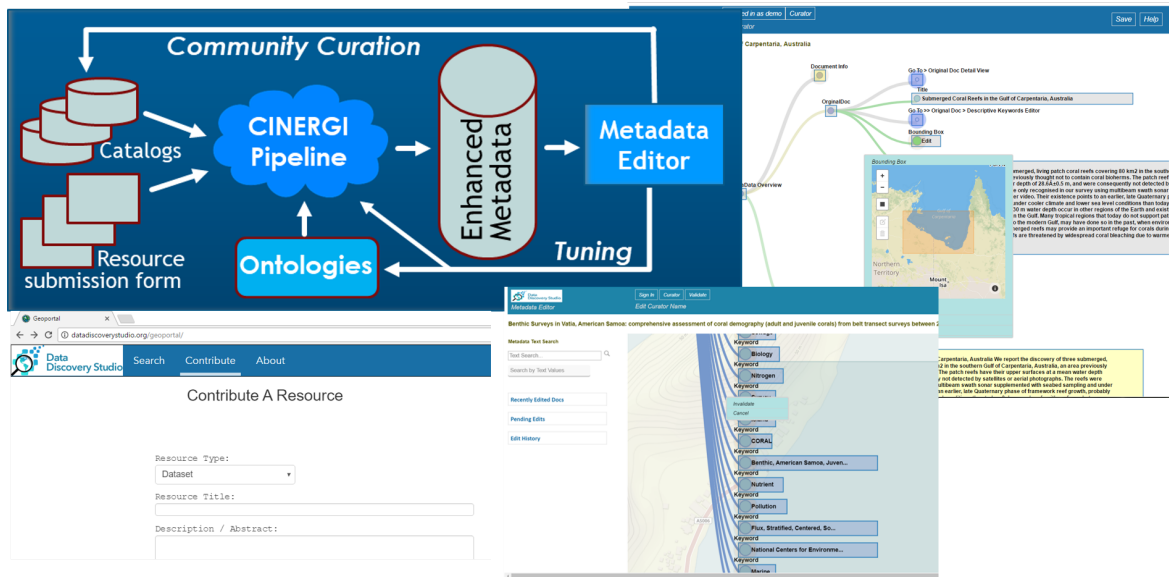
Added annotation properties for combining ontology fragments
(*cinergiFacet*, *cinergiParent*)



Metadata editor
Approve or discard semantic annotations

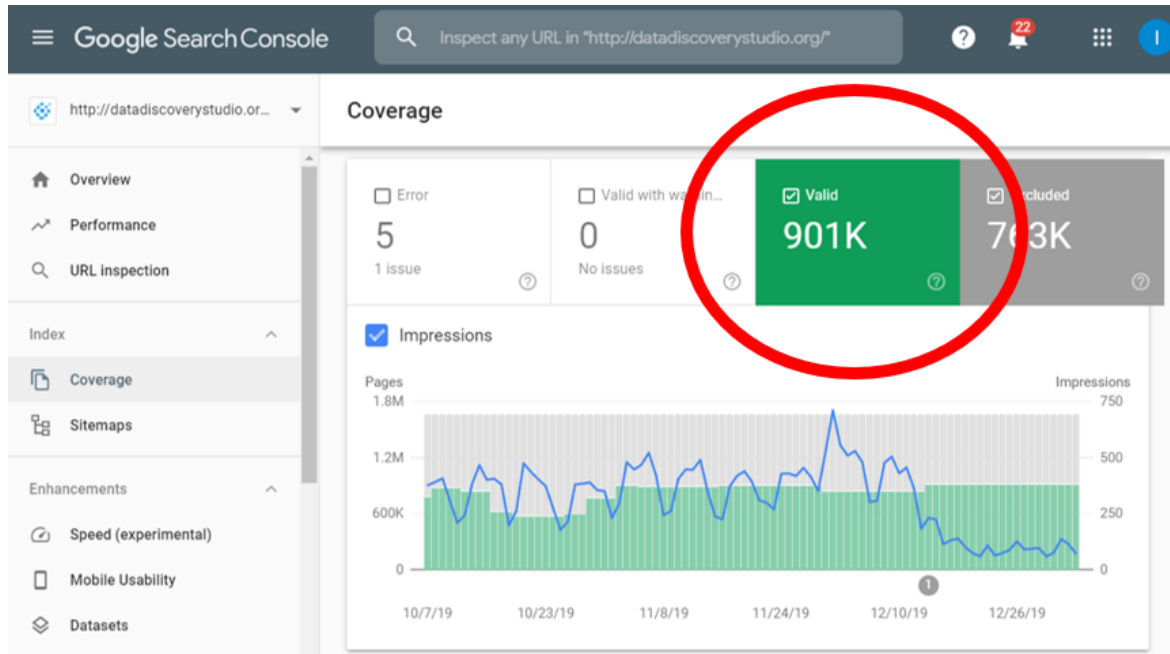


DDStudio experience suggests a scalable metadata curation model, which does not sacrifice domain semantics when improving metadata for discovery. In this model, the first curation step is performed by the automated metadata augmentation pipeline. Then a repository curator can examine the results, invalidating incorrect assignments. This review triggers ontology updates and re-processing, at the same time creating a labeled training set for supervised learning.

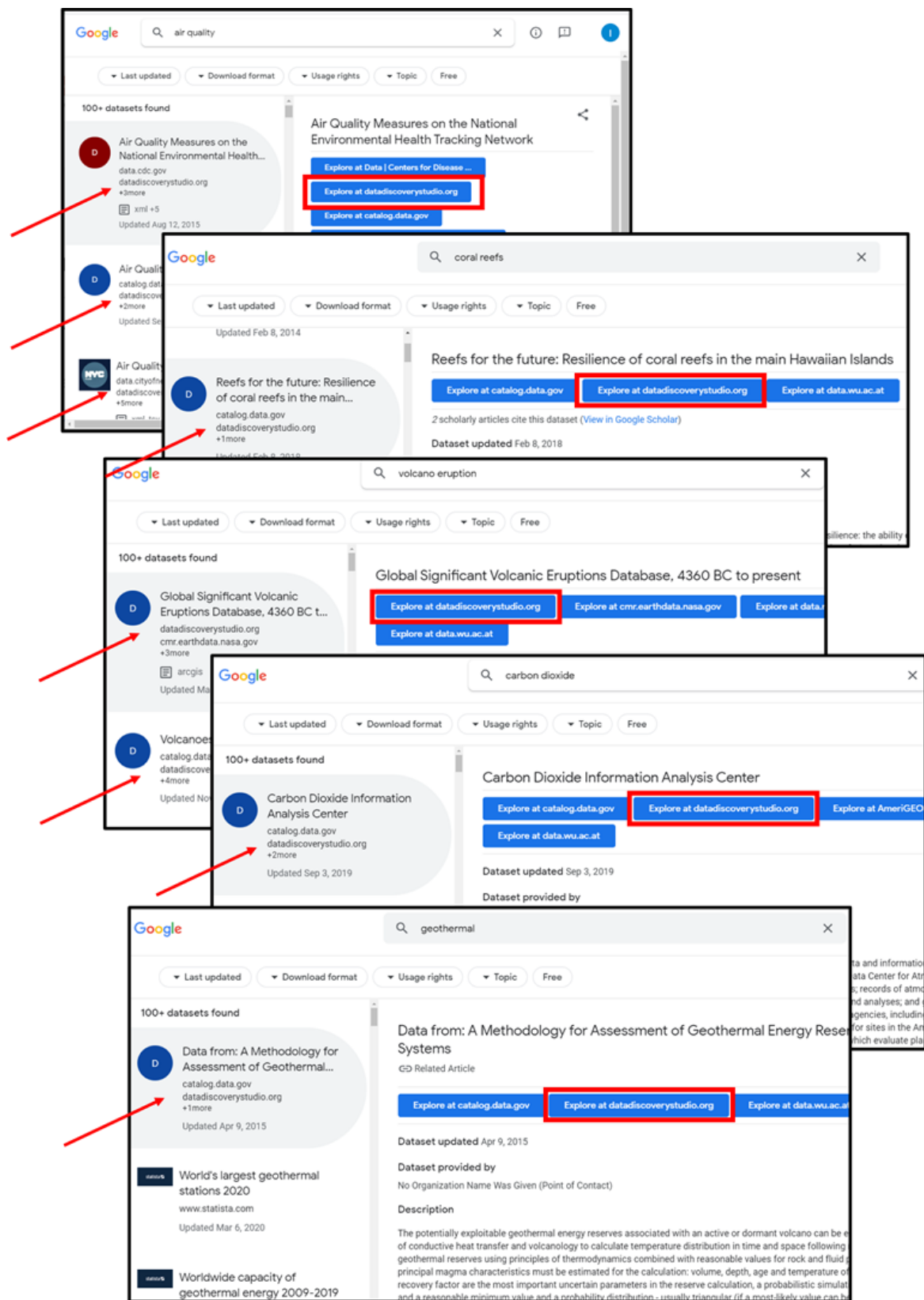


SCHEMA.ORG MARKUP

All metadata records in DDStudio are presented using schema.org markup and referenced in sitemaps submitted to Google. By now, Google indexed 900K records from DDStudio.

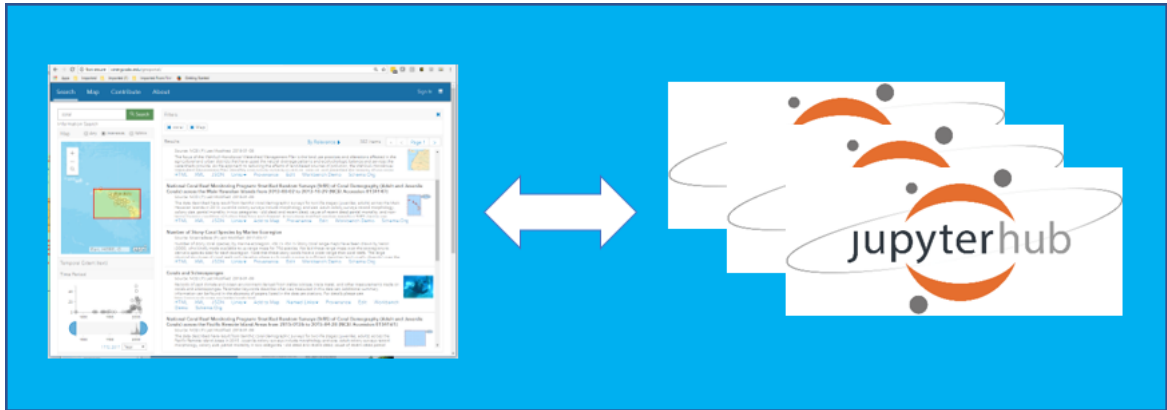


As a result, it is likely that you will be referred to DDStudio when you enter geoscience terms in Google Dataset Search. Please try it yourself!

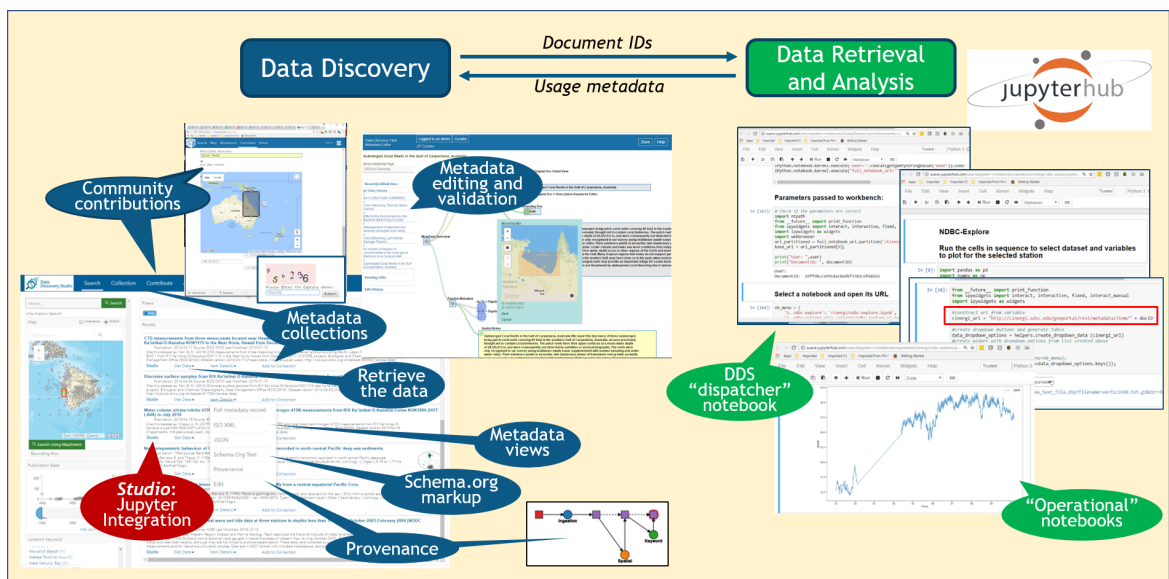


About 40% of DDStudio users have visited the platform via Google referrals.

JUPYTER INTEGRATION



DDStudio prototyped workflows for a seamless transition from geoscience data discovery to research. Using the "Studio" link from search results in the discovery interface, users can launch Jupyter notebooks residing on several JupyterHub servers. The initial link points to a "dispatcher" notebook, which calls other notebooks that implement visualizations, analytical techniques, or models.



The DDStudio system and the collection of Jupyter notebooks are loosely coupled and communicate via standard protocols, which can be used to integrate arbitrary discovery systems and research workbenches using a lightweight interface.

COLLECTION MANAGEMENT

DDStudio implemented a custom version of ESRI Geoportal Server, with the added capability that lets users save any discovered metadata records into collections. The found records can be added to new or existing collection. Users can also export collections and share them with collaborators, or import collections developed by others.

The screenshot displays the DDStudio web application's 'Collection' management section. The top navigation bar includes 'Data Discovery Studio', 'Search', 'Collection' (active), 'Contribute', 'About', 'Tour', and 'Help'. The left sidebar contains a 'Saved Collections' panel with 'Unassigned Results' and a 'Show All' link, and a 'Collection Import' section with 'New Collection', 'Export Collection', 'Collection Info (Merge or Overwrite)', 'Merge', 'Select a CSV File', and 'Import File' buttons. The main content area, titled 'Collection Items', shows a list of three items. The first item, 'A review of the Grenville orogen in its North American type area', includes a detailed text description of the orogen's extent and geological context. The second item, 'Generalized Geology of Europe including Turkey (geo4_2l)', describes the coverage of Europe and its geological features. The third item, 'Grenville-age belts and associated older terranes in Australia and Antarctica', discusses the timing and effects of magmatic and tectonic events. Each item has 'Add to Collection', 'Remove from Collection', and 'Remove Saved record' buttons. At the bottom right, there are links to 'Send Page to a Studio' and 'Send Collection to a Studio'.

Users can also launch Jupyter notebooks for collections. For example, a notebook implementing a landscape model and requiring multiple inputs that exist as separate datasets in DDStudio, can be launched from a collection that organizes such model inputs.

WATCH DATADISCOVERYSTUDIO.ORG IN ACTION!

[VIDEO] <https://www.youtube.com/embed/3opK1o8LgkI?feature=oembed&fs=1&modestbranding=1&rel=0&showinfo=0>

SGCI COLLABORATION

The Science Gateways Community Institute (SGCI) has served as an important partner to DDStudio over the last year.



Find the data you need

Data Discovery Studio, an NSF EarthCube Project, offers a large inventory of high-quality earth science resources with documentation that enables data discovery and re-use.

With over 1.4 million resources in our inventory, you'll have access to data from a wide variety of NSF-supported, government and international scientific repositories and catalogs at no cost to you.

 **Data Discovery Studio**
datadiscoverystudio.org

Key results of our collaboration with SGCI:

- Development of marketing and branding strategies, materials for distribution at professional meetings, and improved social media strategies
- User interface improvements, based on design advice from the SGCI usability team, who conducted user interviews, performed usability testing, and analyzed a dozen of other search portals to identify the most useful features.
- Cybersecurity assessment and single sign-on improvements.