Developing a Scientific Data Search Engine

Part 1: Architecture and Technologies

TEAGAN JOHNSON[1],
Sabira Duishebaeva[1], *Nathan Hook*[2],
Saquib Aziz-Khan[2], *Eric Nienhouse*[2]

SIParCS Intern [1] - NCAR Mentor [2]

July 27, 2022

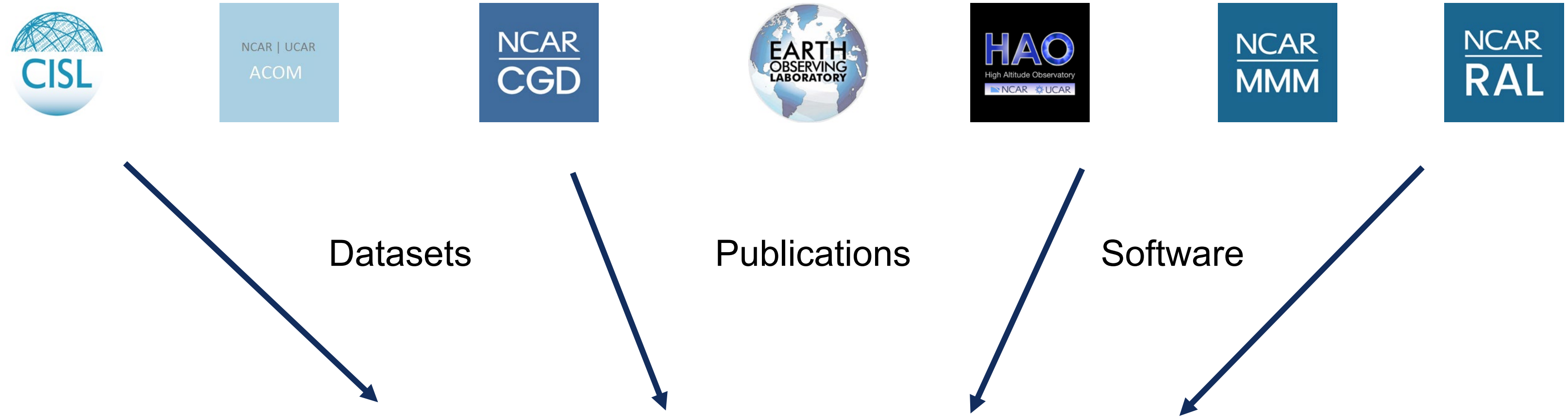
This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Table of Contents

- Background
- Goals
- Methodologies + Technologies
- Architecture
- Result 1: Efficient Deletion
- Acknowledgements

BACKGROUND

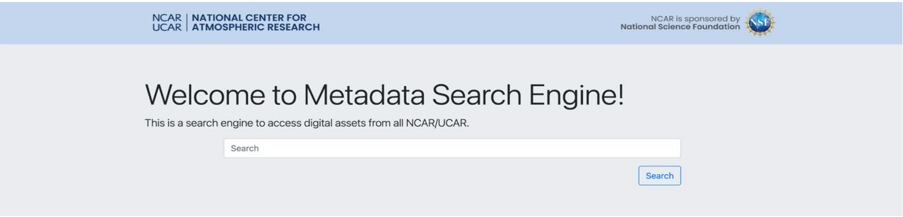
NCAR Labs’ Resources


CISL
NCAR | UCAR ACOM
NCAR CGD
EARTH OBSERVING LABORATORY
HAO High Altitude Observatory NCAR UCAR
NCAR MMM
NCAR RAL

Datasets

Publications

Software


NCAR | NATIONAL CENTER FOR
UCAR | ATMOSPHERIC RESEARCH
NCAR is sponsored by National Science Foundation

Welcome to Metadata Search Engine!

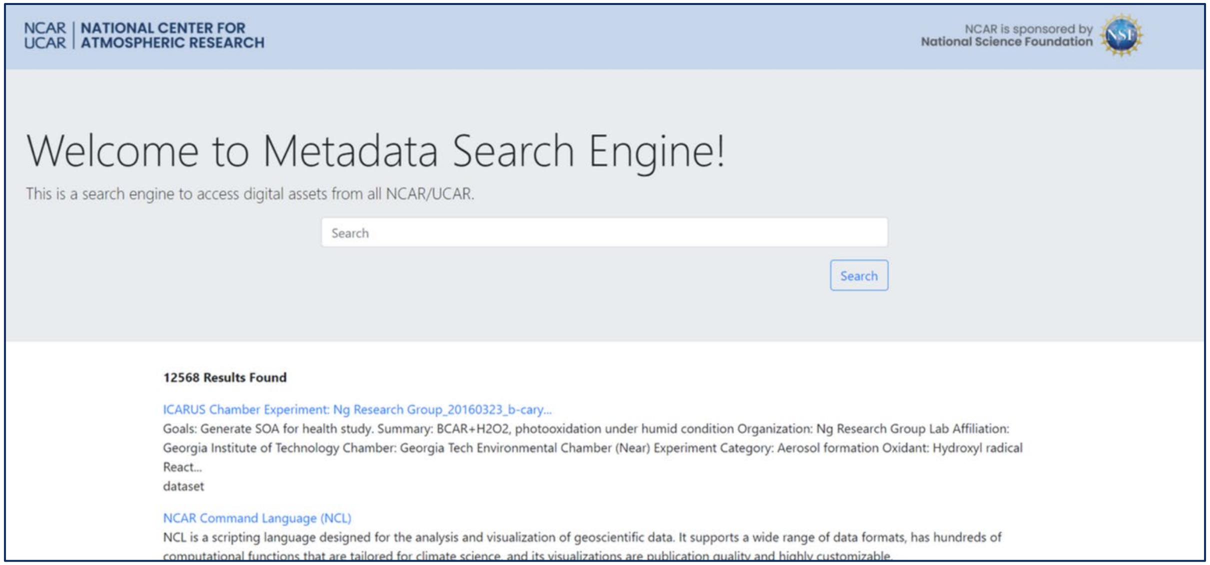
This is a search engine to access digital assets from all NCAR/UCAR.

Where can I find these resources?

General Earth Science Community

Initial State

- Java-based search engine that contained a **harvester** and a **search** application
- Numerous features and bugs needed to be addressed **before deployment**


NCAR | NATIONAL CENTER FOR
UCAR | ATMOSPHERIC RESEARCH
NCAR is sponsored by
National Science Foundation
Search
Search

Welcome to Metadata Search Engine!

This is a search engine to access digital assets from all NCAR/UCAR.

12568 Results Found

ICARUS Chamber Experiment: Ng Research Group_20160323_b-cary...

Goals: Generate SOA for health study. Summary: BCAR+H2O2, photooxidation under humid condition Organization: Ng Research Group Lab Affiliation: Georgia Institute of Technology Chamber: Georgia Tech Environmental Chamber (Near) Experiment Category: Aerosol formation Oxidant: Hydroxyl radical React...

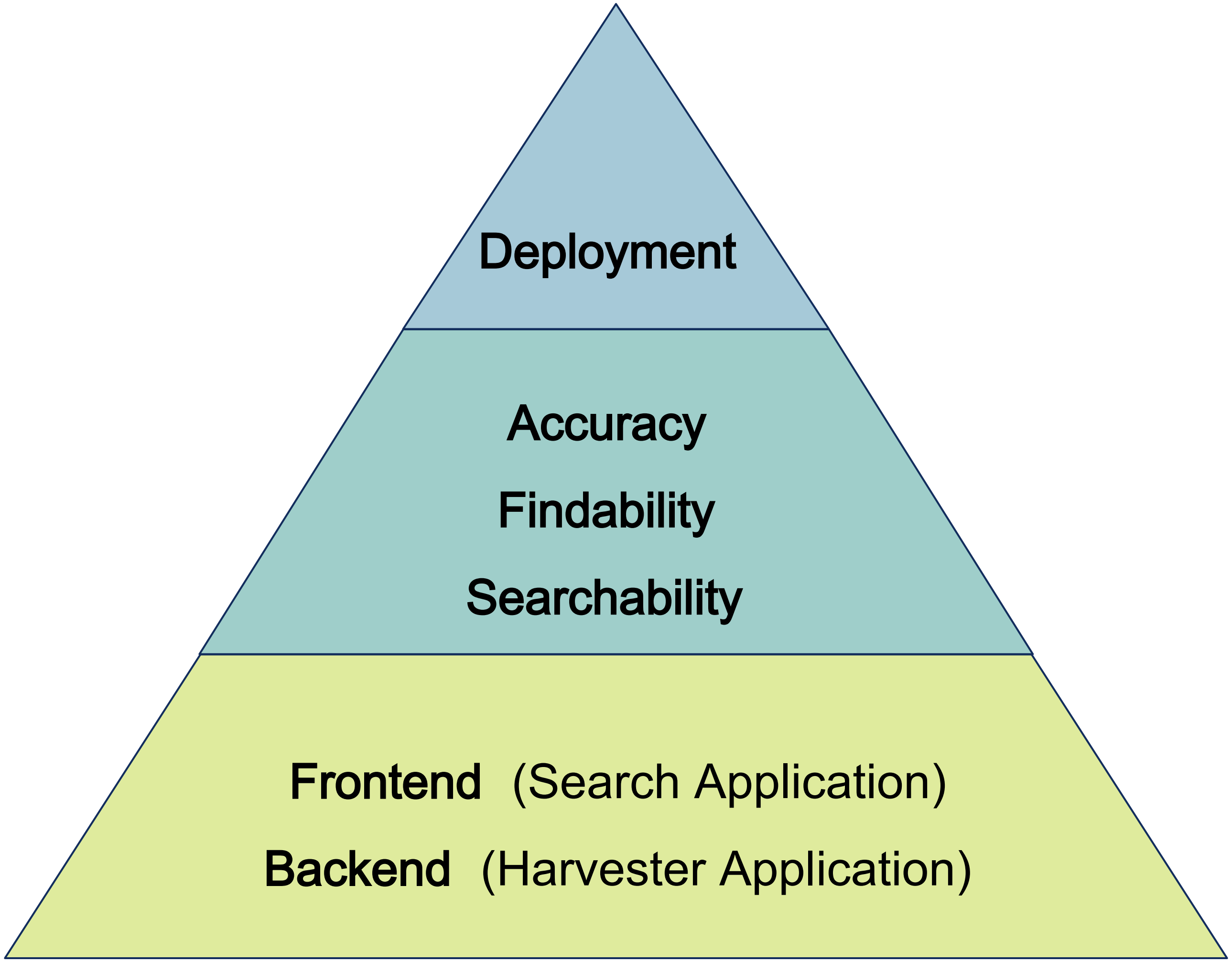
dataset

NCAR Command Language (NCL)

NCL is a scripting language designed for the analysis and visualization of geoscientific data. It supports a wide range of data formats, has hundreds of computational functions that are tailored for climate science, and its visualizations are publication quality and highly customizable.

GOALS

Goals



Deployment

Accuracy

Findability

Searchability

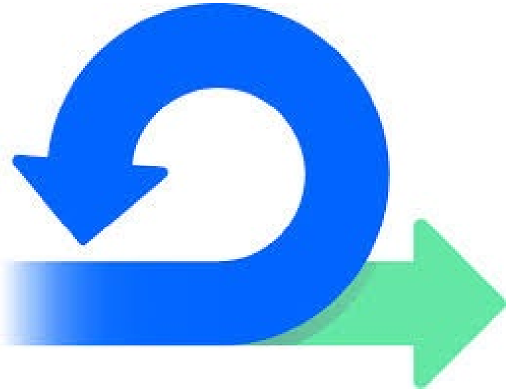
Frontend (Search Application)

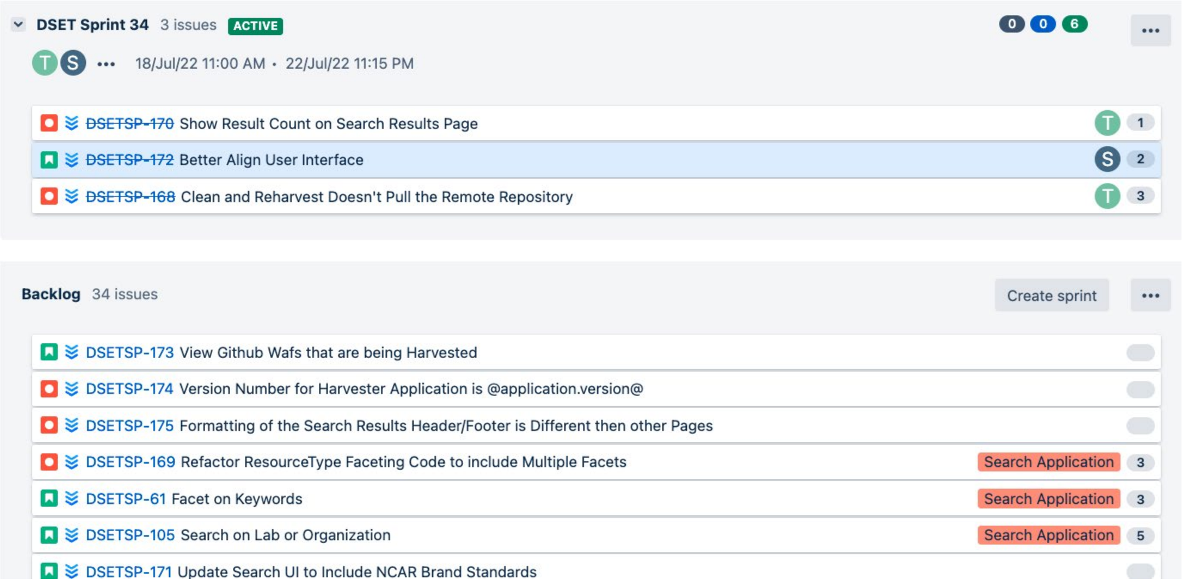
Backend (Harvester Application)

METHODOLOGIES

Agile Scrum

- **Short, iterative cycles** (1 week)
- **Reprioritization** (retro meetings)
- **Communication**





DSET Sprint 34 3 issues ACTIVE 0 0 6 •••

T S ••• 18/Jul/22 11:00 AM • 22/Jul/22 11:15 PM

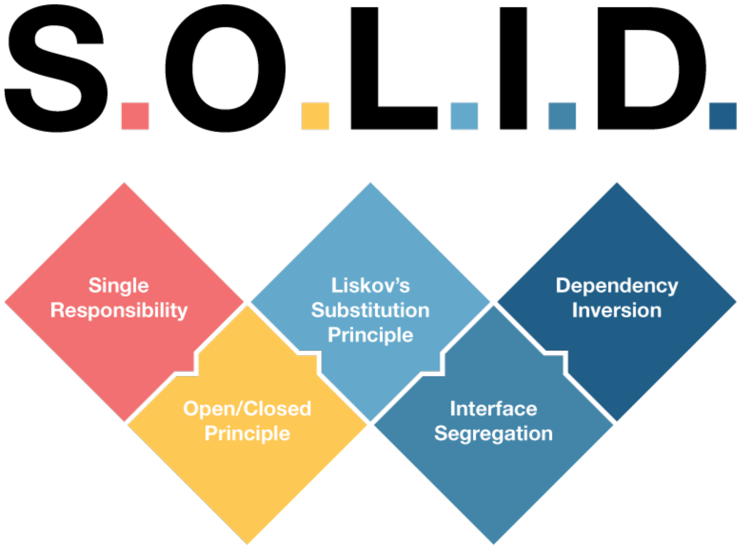
- ~~DSETSP-170~~ Show Result Count on Search Results Page T 1
- ~~DSETSP-172~~ Better Align User Interface S 2
- ~~DSETSP-168~~ Clean and Reharvest Doesn't Pull the Remote Repository T 3

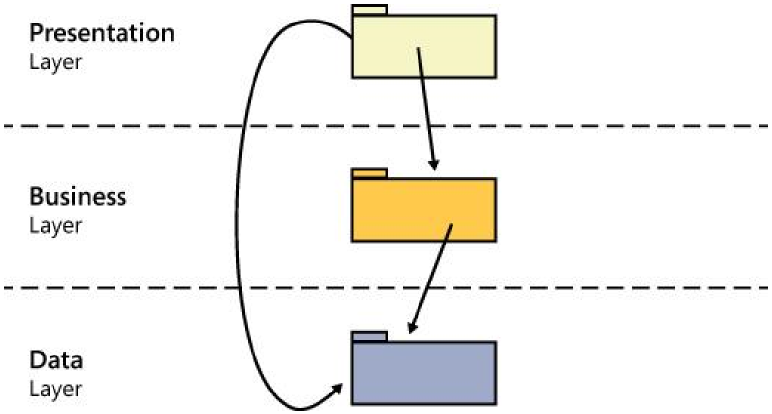
Backlog 34 issues Create sprint •••

- DSETSP-173 View Github Wafs that are being Harvested
- DSETSP-174 Version Number for Harvester Application is @application.version@
- DSETSP-175 Formatting of the Search Results Header/Footer is Different then other Pages
- DSETSP-169 Refactor ResourceType Faceting Code to include Multiple Facets Search Application 3
- DSETSP-61 Facet on Keywords Search Application 3
- DSETSP-105 Search on Lab or Organization Search Application 5
- DSETSP-171 Update Search UI to Include NCAR Brand Standards

Design Principles

- SOLID Principles
- Layered Architecture

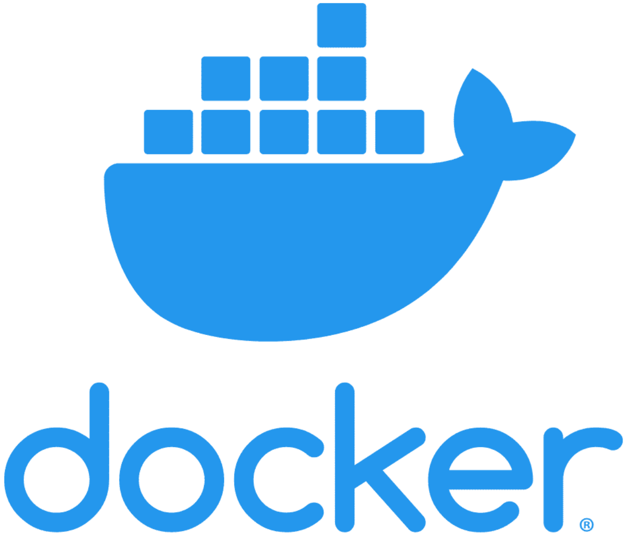




TECHNOLOGIES

spring
boot

Solr


docker

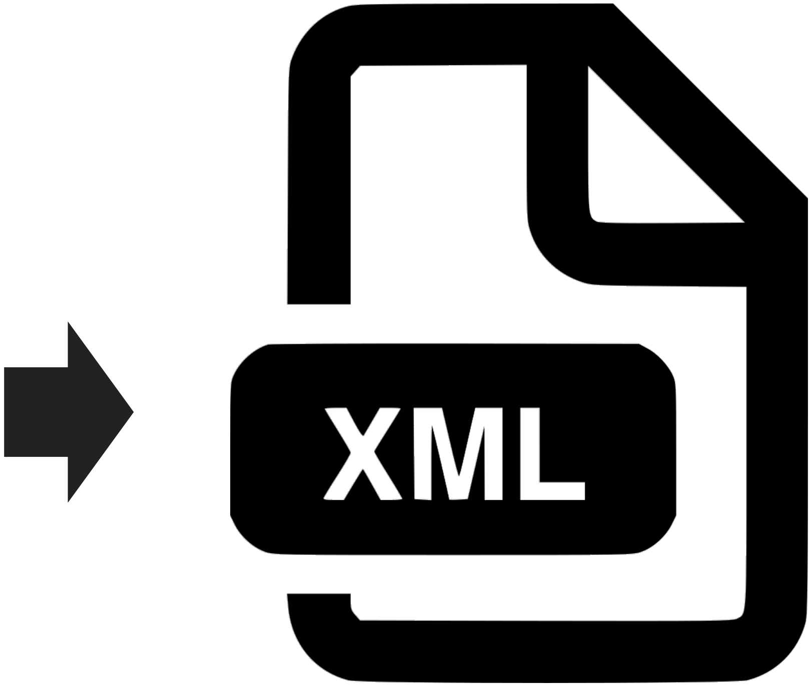
ARCHITECTURE

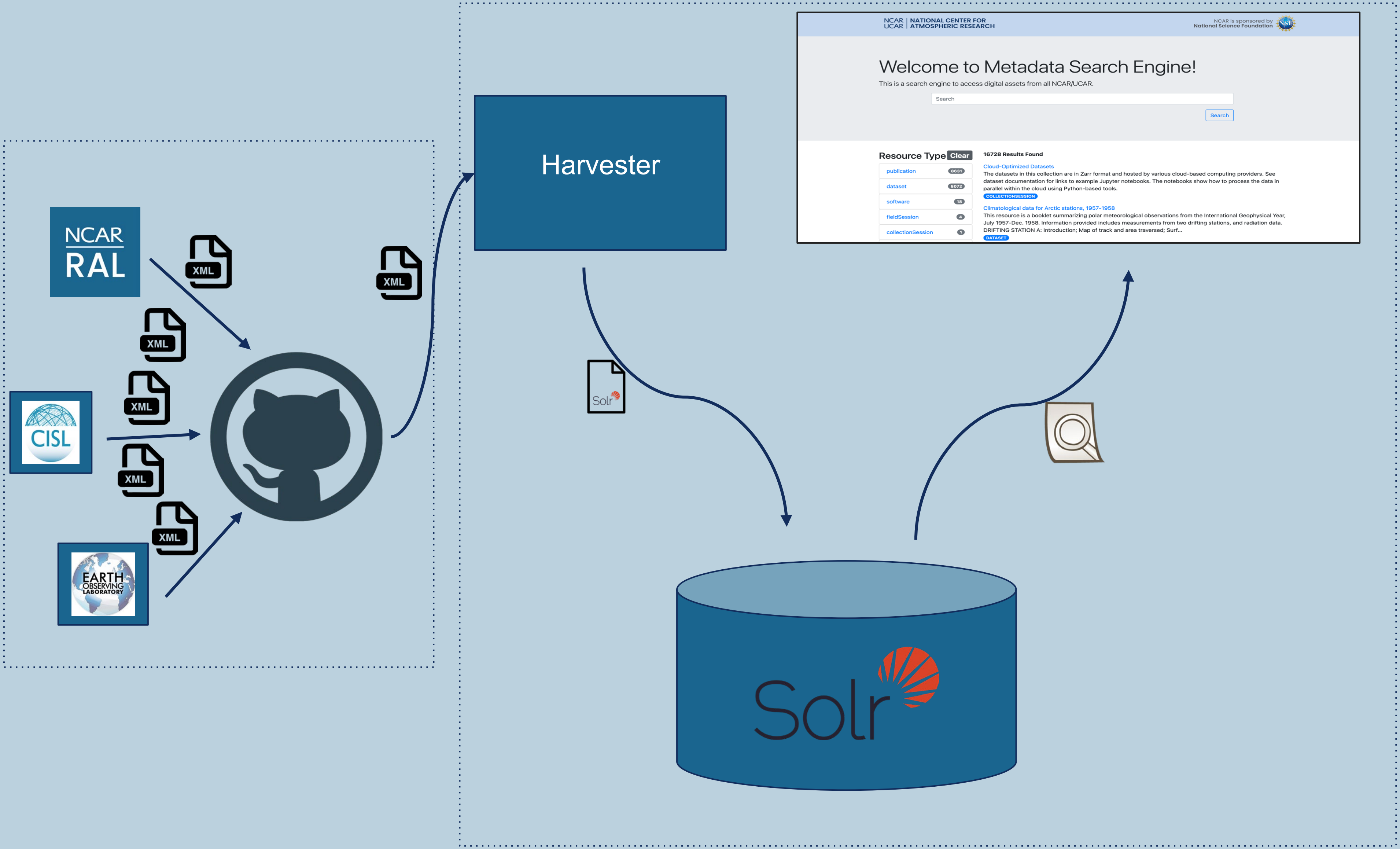
Resource Representation

Every labs’ resources are converted into XML files because XML files are uniform

Properties:

- **Title**
- **ID**
- **Author(s)**
- **Resource Type** (dataset, software, publication, etc.)
- **Lab**
- **Etc.**





NCAR UCAR | NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NCAR is sponsored by National Science Foundation

Welcome to Metadata Search Engine!

This is a search engine to access digital assets from all NCAR/UCAR.

Search

Search

Resource Type Clear

publication	8631
dataset	8072
software	18
fieldSession	4
collectionSession	1

16728 Results Found

Cloud-Optimized Datasets

The datasets in this collection are in Zarr format and hosted by various cloud-based computing providers. See dataset documentation for links to example Jupyter notebooks. The notebooks show how to process the data in parallel within the cloud using Python-based tools.

COLLECTIONSESSION

Climatological data for Arctic stations, 1957-1958

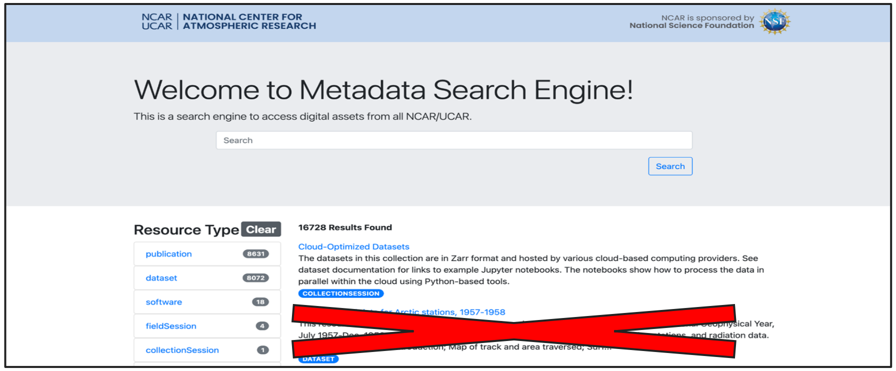
This resource is a booklet summarizing polar meteorological observations from the International Geophysical Year, July 1957-Dec. 1958. Information provided includes measurements from two drifting stations, and radiation data. DRIFTING STATION A: Introduction; Map of track and area traversed; Surf...

DATASET

RESULT 1: DELETION

The Problem

- If a lab deletes a file, it should be deleted from the search engine **(increase accuracy)**



NCAR | NATIONAL CENTER FOR
UCAR | ATMOSPHERIC RESEARCH

NCAR is sponsored by
National Science Foundation

Welcome to Metadata Search Engine!

This is a search engine to access digital assets from all NCAR/UCAR.

Search

Search

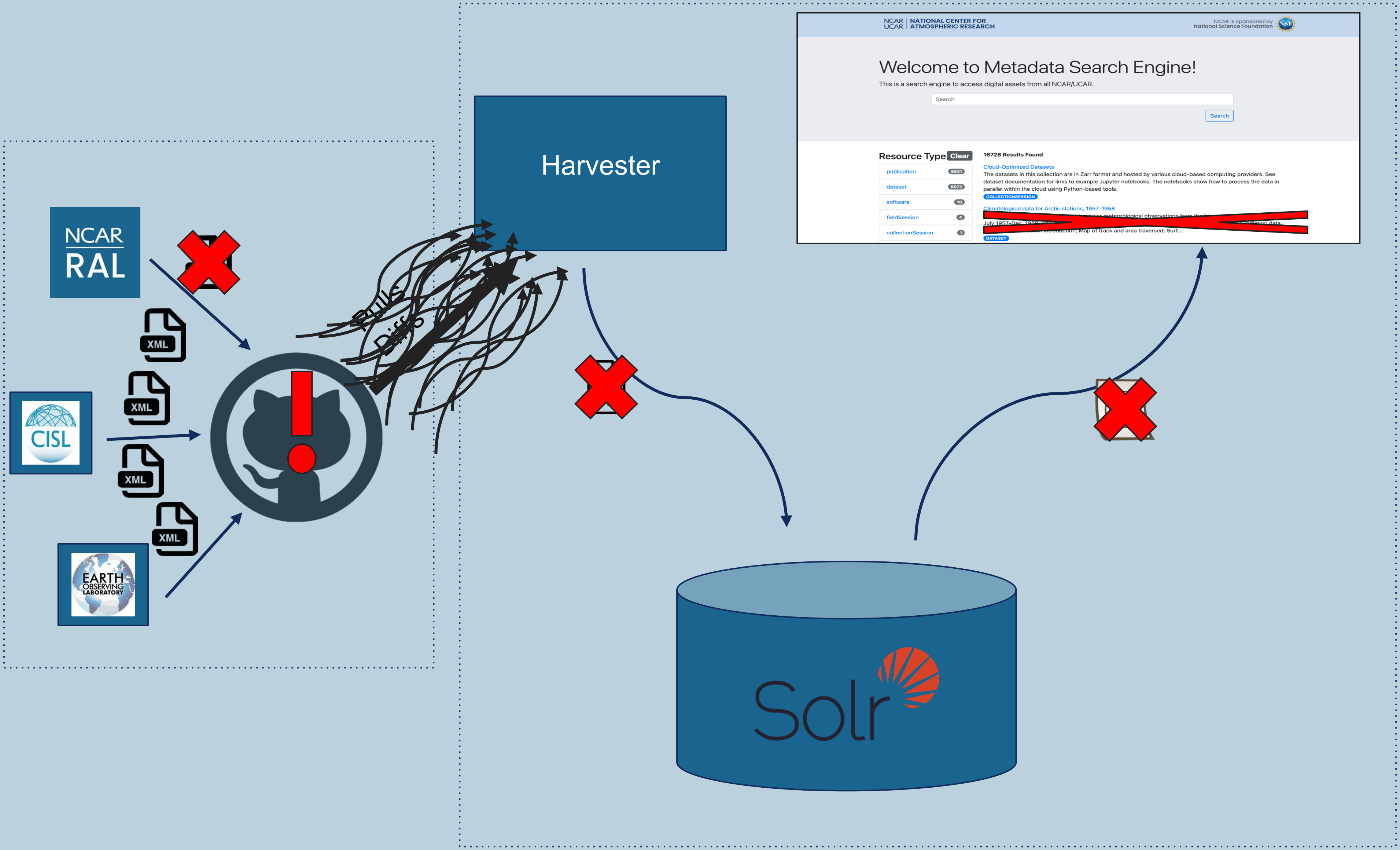
Resource Type Clear	
publication	8631
dataset	8072
software	18
fieldSession	4
collectionSession	1

16728 Results Found

Cloud-Optimized Datasets
The datasets in this collection are in Zarr format and hosted by various cloud-based computing providers. See dataset documentation for links to example Jupyter notebooks. The notebooks show how to process the data in parallel within the cloud using Python-based tools.
COLLECTIONSESSION

~~for Arctic stations, 1957-1958~~
~~This resea… Geophysical Year, July 1957-Dec 195… ations, and radiation data. …oduction; Map of track and area traversed; Sum...~~
DATASET

- Before, to delete a file we'd clear the harvester and re -pull *every* file from Github which was **very inefficient**



NCAR | UCAR | NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NCAR is sponsored by National Science Foundation

Welcome to Metadata Search Engine!

This is a search engine to access digital assets from all NCAR/UCAR.

Search

Search

Resource Type Clear

publication	8631
dataset	8072
software	18
fieldSession	4
collectionSession	1

16728 Results Found

Cloud-Optimized Datasets

The datasets in this collection are in Zarr format and hosted by various cloud-based computing providers. See dataset documentation for links to example Jupyter notebooks. The notebooks show how to process the data in parallel within the cloud using Python-based tools.

COLLECTIONSESSION

Climatological data for Arctic stations, 1957-1958

...polar meteorological observations from the Int...

July 1957-Dec. 1958 ... radiation data.

... Introduction; Map of track and area traversed; Surf...

DATASET

Efficient Deletion

- To delete one file it took **11 mins, 36 secs**
- Now it takes **LESS THAN A SECOND**

99%

ACKNOWLEDGEMENTS

Thank You

NCAR and CISL

SIParCS Mentors:

- Nathan Hook
- Saquib Aziz Khan
- Eric Nienhouse

Project Partner:

- Sabira Duishebaeva

SIParCS Program Leads:

- Virginia Do
- Jerry Cyccone
- AJ Lauer
- Francesgladys Pulido

And everyone else involved with the SIParCS program this summer.

Resources

- https://www.hiclipart.com/ - Most images in presentation
- https://blog.knoldus.com/why -we-need-solid-principles-and-its-types/ - SOLID Principles
- https://www.youtube.com/c/ncarcgd - CGD logo
- https://mobile.twitter.com/ncar_acom - ACOM, EOL logos
- https://mobile.twitter.com/ncar_cisl CISL logo
- https://www.youtube.com/channel/UCgDMGLn6JKEU87aJJlt8F-g - HAO logo
- https://mobile.twitter.com/ncar_mmm - MMM logo
- https://www.youtube.com/channel/UCDIFPOu2f7TortTgNYQc6wA - RAL logo
- https://www.scrum.org/resources/scrum -framework-poster - Scrum Diagram