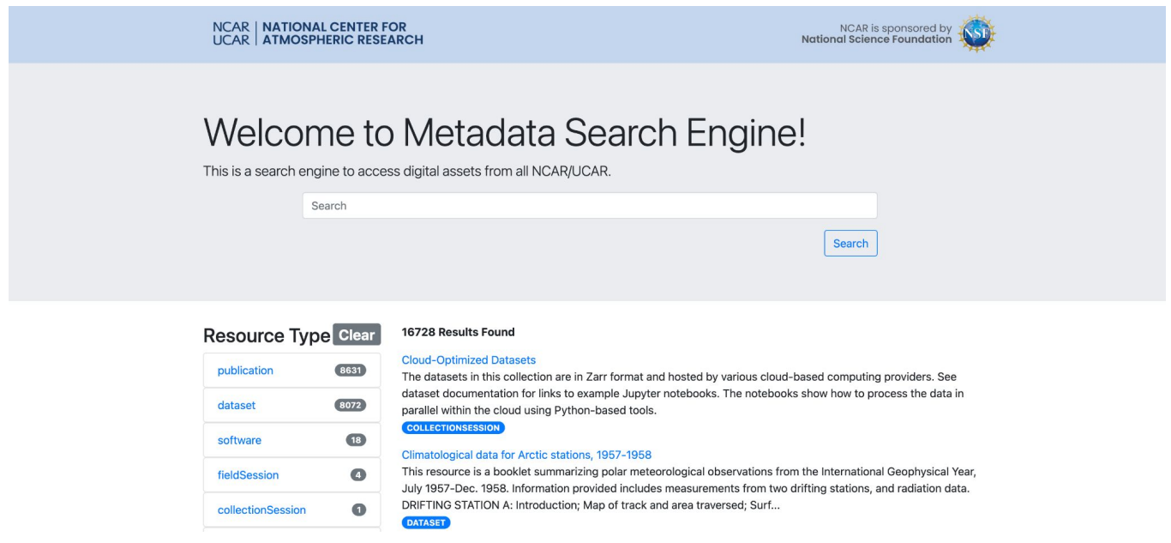
DEVELOPING A SCIENTIFIC DATA SEARCH ENGINE

Teagan Johnson, Sabira Duishebaeva
Mentors: Nathan Hook, Saquib Aziz-Khan, Eric Nienhouse

INTRO

NCAR's Search Engine

NCAR has a variety of research labs that produce resources such as datasets, publications, and software. It's important that these resources are accessible for end users which is why NCAR has a search engine. This summer, **we worked on a new search engine** that has been in development for the past two years.



NCAR | NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
UCAR
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

Metadata

In the context of the search engine, we refer to these resources as **metadata**.

The "Problem"

Eventually, this new search engine may be deployed and be used as NCAR's primary search engine. Before it's deployed, there are many features and bugs that need to be addressed. Our job this summer was to **push the search engine closer to deployment.**

OUR WORK

Validated Metadata

Implemented a validation feature for the scientific metadata to **ensure search results are accurate.**

Improved Deletion Efficiency

Decreased the amount of time it takes to delete a file by up to 100%, **further improving the search engine's accuracy.**

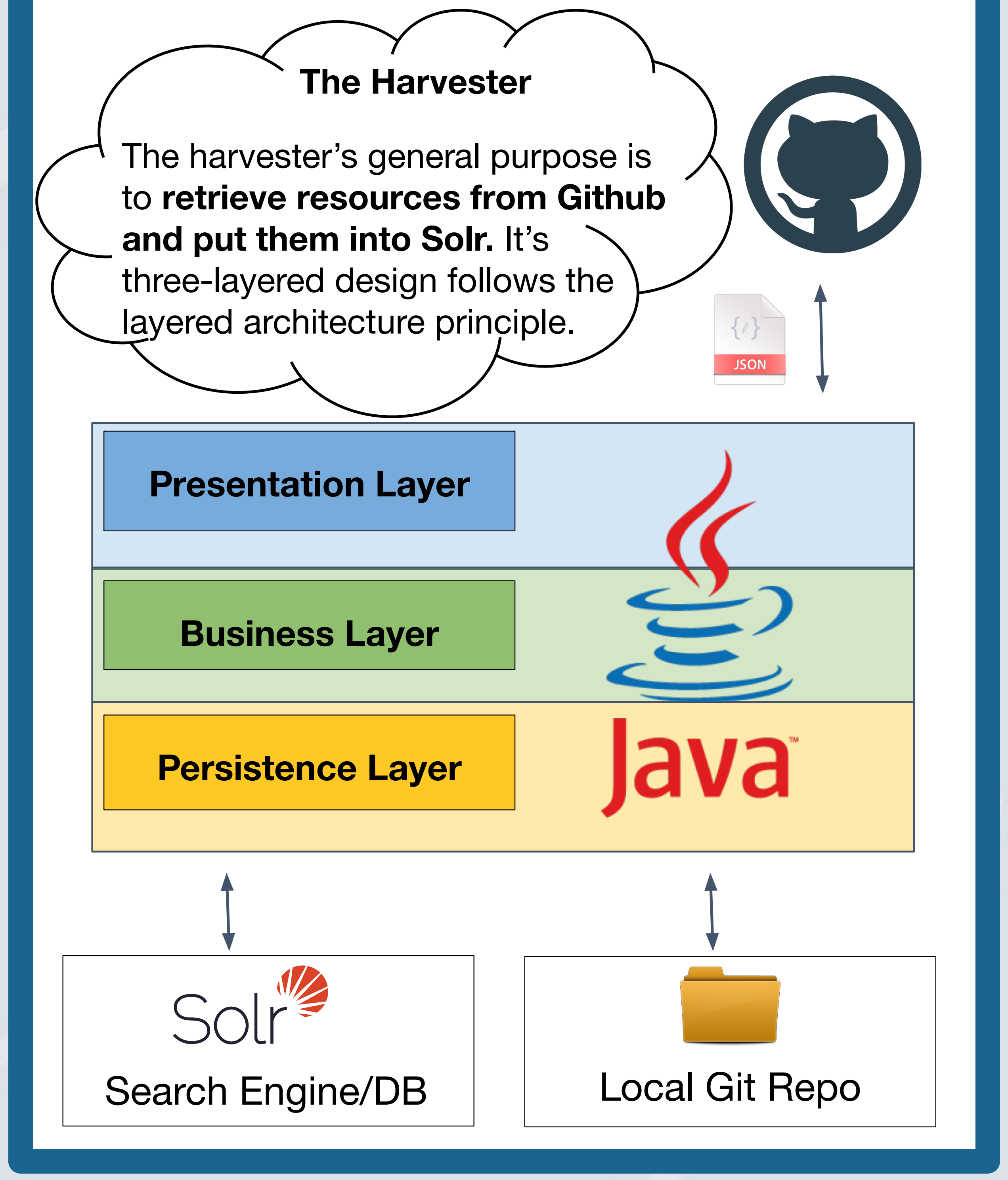
Enabled Search Faceting

Designed a user-facing facet feature that enables filtering search results by various criteria, **improving the search engine's searchability.**

Implemented Google Indexing

Provided a way for Google to efficiently crawl and index results in the search engine with a sitemap and JSON-ld, **revamping the search engine's findability.**

HARVESTER ARCHITECTURE



The Harvester

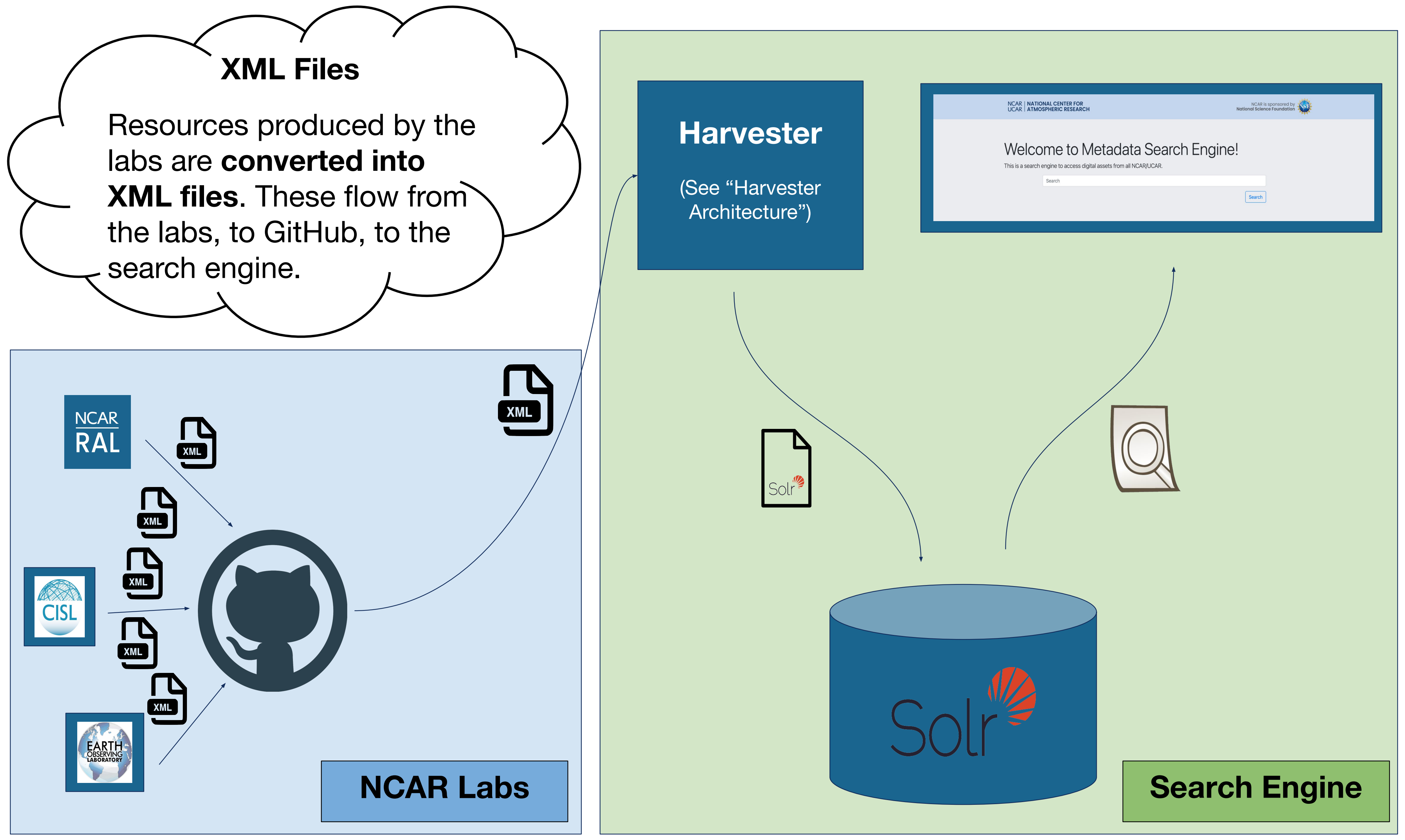
The harvester's general purpose is to **retrieve resources from Github and put them into Solr.** It's three-layered design follows the layered architecture principle.

SOFTWARE

DATA FLOW

XML Files

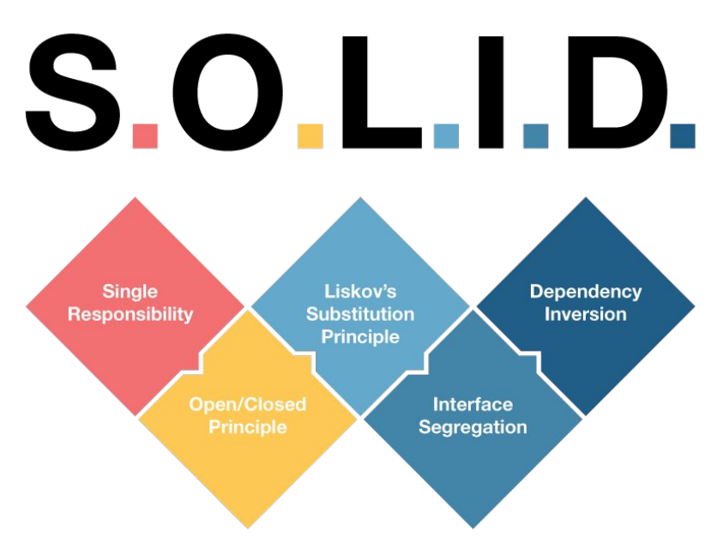
Resources produced by the labs are **converted into XML files**. These flow from the labs, to GitHub, to the search engine.



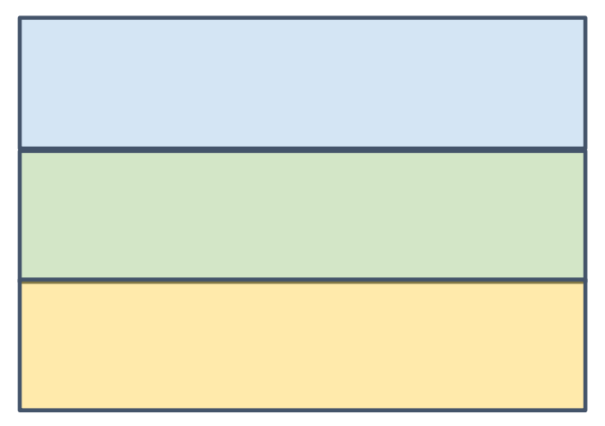
Harvester
(See "Harvester Architecture")
Welcome to Metadata Search Engine!
NCAR RAL
CISL
EARTH OBSERVING LABORATORY
XML
Solr
NCAR Labs
Search Engine

METHODOLOGIES

Scrum



Layered Architecture



CONCLUSION

Conclusion

We ultimately achieved our goal of progressing the search engine towards deployment by **improving its searchability, findability, and accuracy.** More specifically, we implemented metadata validation, search faceting, efficient deletion, and Google indexing.

Future Work

Future work includes adding extensions to validation and faceting, implementing autofill and spellcheck algorithms, designing a login feature for the harvester, and more.

ACKNOWLEDGMENTS

Thank you to our mentors **Nathan Hook**, **Saquib Aziz-Khan**, and **Eric Nienhouse**, to the SIParCS coordinators **Virginia Do**, **AJ Lauer**, **Jerry Cyconne**, and **FrancesGladys Pulido**, and to the NSF for this project and to NCAR and CISL for their support of SIParCS.