DEVELOPING A SCIENTIFIC DATA SEARCH ENGINE

Sabira Duishebaeva, Teagan Johnson

Mentors: Nathan Hook, Saquib Aziz-Khan, Eric Nienhouse

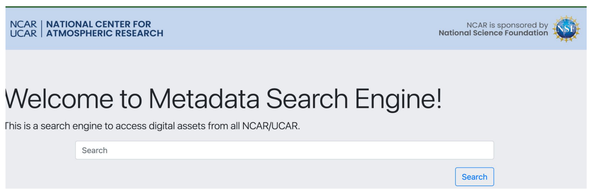
INTRO

NCAR's Search Engine

NCAR's diverse scientific data holdings have historically been difficult for external scientists and users to search across and find the data they need to do their science. While we have a current search system that aggregates these data holdings, **we are experimenting with a simpler approach.**

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.**

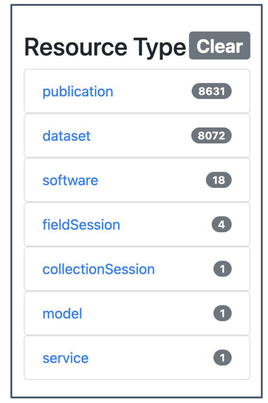
Implemented Google Indexing

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.**

SEARCH FACETING

Why need?

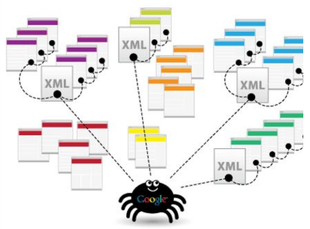
Data providers categorize their files by r**esource type.** We want to be able to **filter** metadata search results by their resource type.



Future Work

- Adding **multi-select** options
- Adding more options to facet by such as **labs, keywords, authors** etc.

SITEMAP/JSON-LD



Why need?

A good XML sitemap acts as a roadmap of your website that leads Google to all your important pages and JSON-LD describes it.

JSON-LD on Our Web App

```
{
  "@context" : "http://schema.org",
  "@type" : "Dataset",
  "@id" : "https://doi.org/10.5065/7c2d-bg23",
  "identifier" : "https://doi.org/10.5065/7c2d-bg23",
  "name" : "High Resolution Historical and Future Simulations Over Hawaii",
  "description" : "To better understand the rainfall climatology and its impacts on hy
  "url" : "https://doi.org/10.5065/7c2d-bg23",
```

Future Work

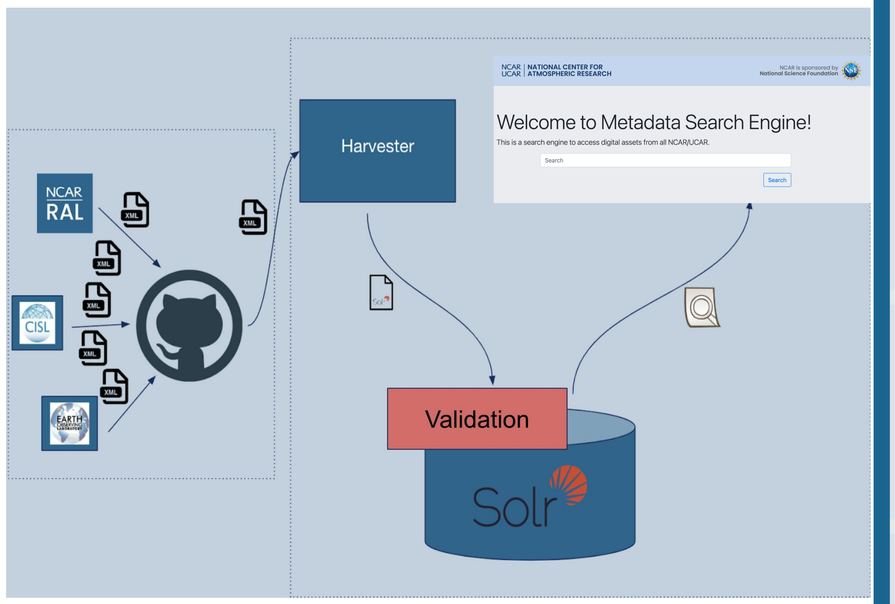
Adding Robots.txt to be able to crawl only **specific pages** of the website.

VALIDATION

H

Results

- Only **valid** records displayed
- More **reliable** and **consistent** data
- Data providers can **view** the error messages



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.