Developing a Scientific Data Search Engine

Part 2: Features & Results

Sabira Duishebaeva [1],
Teagan Johnson[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
Metadata Validation

Feature#1: Metadata Validation

Why need?

- Improves searchibility and findability
- Keeps data consistent & reliable

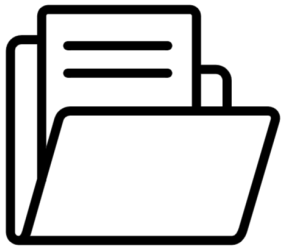
DSET Rules:

Concepts	Definition
- Minimum Required	
- Author	The person(s)/institution(s) receiving credit, as in a citation.
- Title	A name given to the data set, model, software or other asset.
- ISO Asset Type	Type of asset.
- Landing page	Web accessible landing page.
- Publication Date	Date asset was first made available.
- Metadata date	Date stamp when metadata record created or last updated
- Publisher	The lab (or smaller group) that made asset available.
- Resource Support Contact	Person, group, or institution to contact for support on asset.
- Metadata point of Contact	Party responsible for the metadata
- Description	A summary of data set content, or description of asset.

Metadata Validation: How Does It Work?

Data Providers

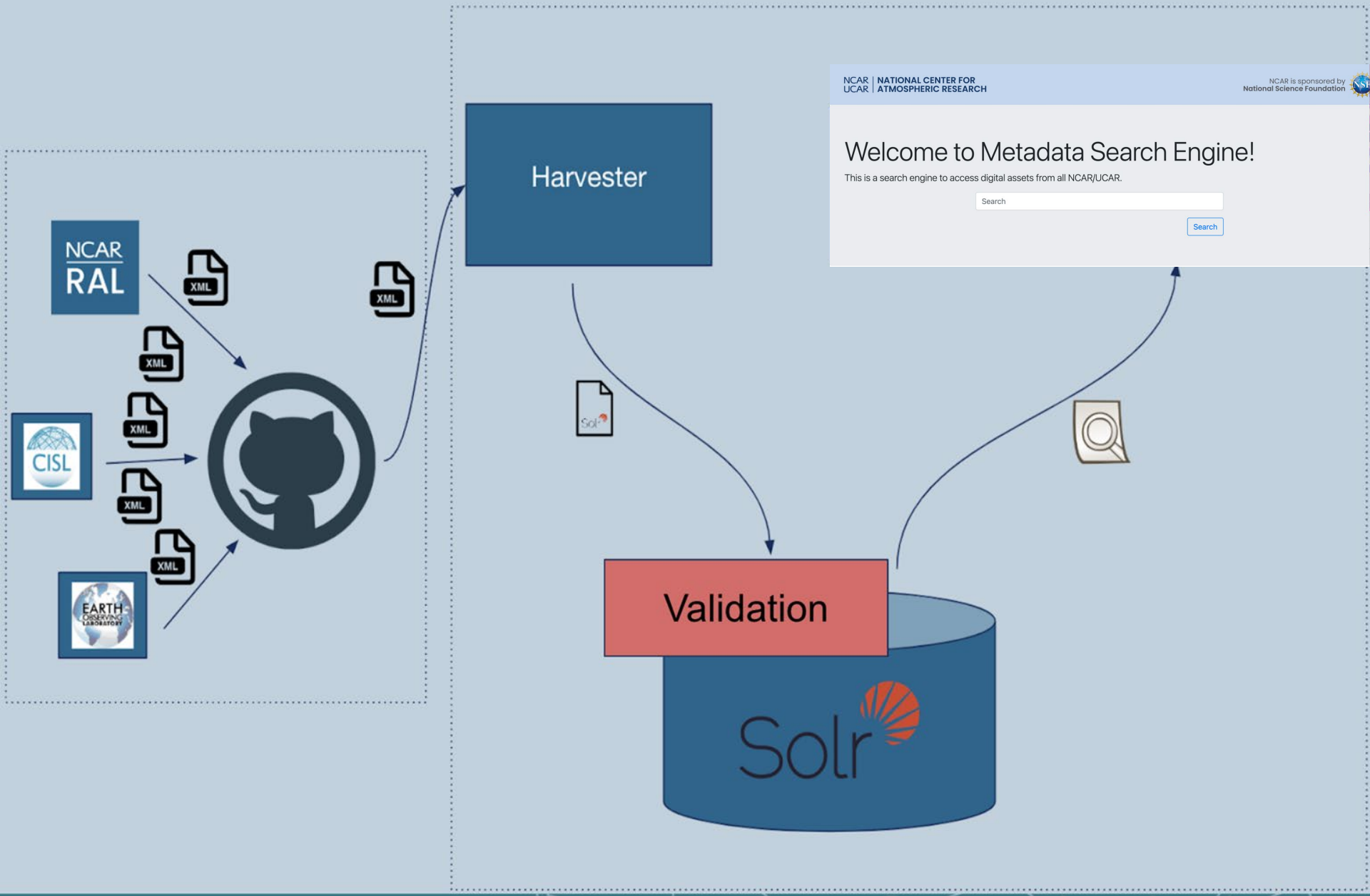
Record their data



Put them in xml files

XML Files in GitHub

```
<?xml version="1.0" encoding="UTF-8" ?>
<gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:gco="http://www.isotc211.org/2005/gco" xmlns:gml="http://www.opengis.net/gml" xmlns:xsi="htt
  <gmd:fileIdentifier>
    <gco:CharacterString>edu.ucar.rda::ds010.0</gco:CharacterString>
  </gmd:fileIdentifier>
  <gmd:language>
    <gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/" codeListValue="eng; USA">eng; USA</gmd:LanguageCode>
  </gmd:language>
  <gmd:characterSet>
    <gmd:MD_CharacterSetCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_CharacterSetCode" codeListValue="utf8">utf8</gmd:MD_C
  </gmd:characterSet>
  <gmd:hierarchyLevel>
    <gmd:MD_ScopeCode codeList="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#MD_ScopeCode" codeListValue="dataset">dataset</gmd:MD_ScopeCode
  </gmd:hierarchyLevel>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:organisationName>
        <gco:CharacterString>NCAR Research Data Archive</gco:CharacterString>
      </gmd:organisationName>
      <gmd:contactInfo>
        <gmd:CI_Contact>
          <gmd:phone>
            <gmd:CI_Telephone>
              <gmd:facsimile>
                <gco:CharacterString>303-497-1291</gco:CharacterString>
              </gmd:facsimile>
            </gmd:CI_Telephone>
          </gmd:phone>
          <gmd:address>
            <gmd:CI_Address>
              <gmd:deliveryPoint>
                <gco:CharacterString>National Center for Atmospheric Research</gco:CharacterString>
              </gmd:deliveryPoint>
              <gmd:deliveryPoint>
                <gco:CharacterString>CISL/DSS</gco:CharacterString>
              </gmd:deliveryPoint>
```


NCAR
RAL
XML
XML
XML
XML
XML
CISL
EARTH
OBSERVING
LABORATORY
Harvester
Solr
Validation
Solr
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

XML Files in Solr


```
{
  "id":"edu.ucar.rda::ds061.0",
  "title":"NMC 47 by 51 Northern Hemisphere Stratospheric Analyses, daily 1964-1980",
  "description":"This dataset of daily gridded analyzed geopotential height and temperatur
  "doi":"https://doi.org/10.5065/Y7MH-0127",
  "keywords":["EARTH SCIENCE > ATMOSPHERE > ALTITUDE > GEOPOTENTIAL HEIGHT"],
  "resource_type":"dataset",
  "authoritative_source_url":"https://doi.org/10.5065/Y7MH-0127",
  "authoritative_source_location_on_disk":"/Users/sduishebaeva/Java/xml/test-pull-method/t
  "authoritative_source_md5":"fa845f9e23cacf2923e89416c11de82f",
  "github_xml_url":"https://github.com/dsabira/test-pull-method.git/blob/main/test11.xml",
  "is_valid":true,
  "index_timestamp":"2022-07-15T17:11:34.690Z",
  "_version_":1738439492569661440},
{
  "id":"10.5065/9n3z-7x72",
  "description":"The PyConform package is a Python-based package for converting model time-
  "doi":"https://doi.org/10.5065/9n3z-7x72",
  "keywords":["Software"],
  "resource_type":"dataset",
  "authors":["Paul, Kevin",
    "Mickelson, Sheri",
    "Dennis, John"],
  "author_emails":["",
    "",
    ""],
  "authoritative_source_url":"https://doi.org/10.5065/9n3z-7x72",
  "authoritative_source_location_on_disk":"/Users/sduishebaeva/Java/xml/test-pull-method/t
  "authoritative_source_md5":"8e5492f804fe8744a516c8a71ba64cbb",
  "github_xml_url":"https://github.com/dsabira/test-pull-method.git/blob/main/test9.xml",
  "is_valid":false,
  "validation_messages":["Title must not be empty"],
  "index_timestamp":"2022-07-15T17:11:36.370Z",
  "_version_":1738439494331269120},
{
```

Results & Future Work
Validation

Results & Future Work for Metadata Validation

Results:

- Consistent & reliable metadata search results
- Error messages in the Solr

Future Work:

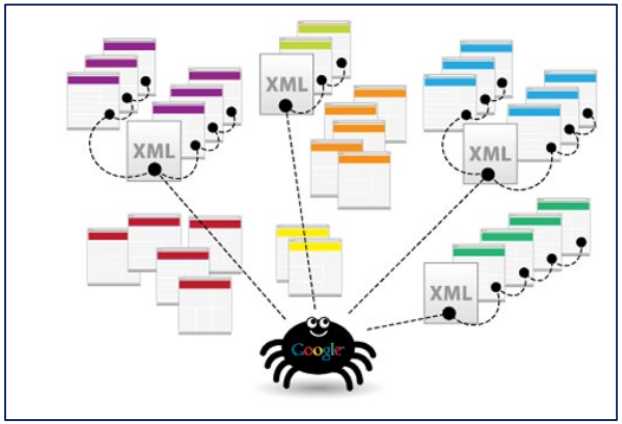
- More complex validation
- Notifications to data providers

BACKGROUND
Sitemap/JSON -LD

Feature#2: Sitemap/JSON -LD

Why need?

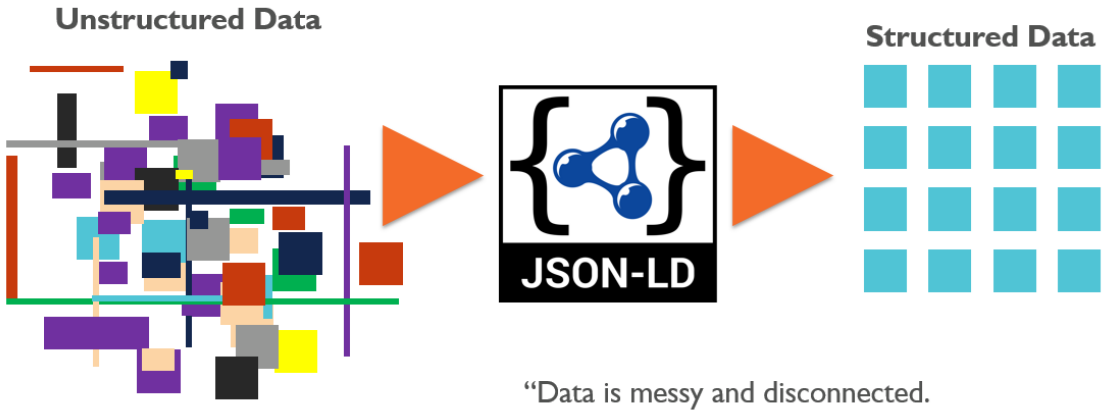
- A good XML sitemap acts as a roadmap of your website that leads Google to all your important pages and JSON-LD describes it.




```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/org.earthsystemgrid.www::0c156aac-ba21-4092-8c9d-0e4e62ebf933</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/f2dfa88e-3abb-11e9-8f53-b808cf016134</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/7f1a2b80-8976-42bb-a2b4-b3d1fbf335c7</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/ca38361d-3aaf-11e9-8f53-b808cf016134</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/edu.ucar.cisl::cloud-collection</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/1f325eee-0d1a-428a-b2f8-6e5eba3652d2</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/dcu-UDI-1958-20150320113628.0-ocm01332775</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/nl-OLA-1968-20150320161613.0-ocm09338459</loc>
    </url>
    <url>
        <loc>http://sagedockerdev.ucar.edu:8080/metadata/id/ncu-GPO-1982-20150313184310.0-ocm10283357</loc>
    </url>
    <url>
```

Feature#2: Sitemap/JSON -LD

What does JSON-LD do?



"Data is messy and disconnected. JSON-LD organizes and connects it, creating a better Web."

Results & Future Work
Sitemap/JSON -LD

Results: JSON -LD on Our Web Application

NCAR | NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
UCAR

NCAR is sponsored by National Science Foundation

High Resolution Historical and Future Simulations Over Hawaii

Id:
org.earthsystemgrid.www::0c156aac-ba21-4092-8c9d-0e4e62ebf933

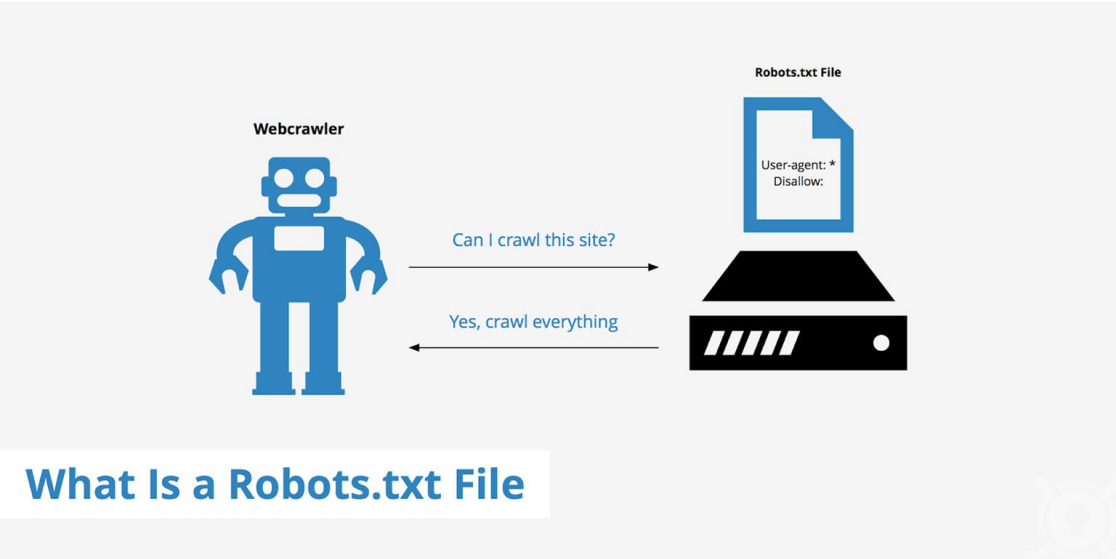
Description:
To better understand the rainfall climatology and its impacts on hydrological cycle over the Hawaiian Islands under historical and future climates, regional climate simulations over the main Hawaiian islands have been conducted for two 10-year periods using the Weather Research and Forecasting (WRF) model in a configuration of two nested domains. The historical 10-year simulation was driven by the ERA-Interim global reanalysis data and observed sea surface temperature from Oct. 2002 to Sep. 2012 (historical simulation). A high-resolution vertical coordinate was employed to better resolve the trade wind inversion (TWI). Results show that the historical simulation reproduces the mean surface temperature, relative humidity and winds with low biases (+/- 1 degree C, +/- 4% and +/- 1 m s-1, respectively) and high spatial correlations (r > 0.80). Additionally, for the historical simulation WRF accurately reproduced aggregated daily and hourly rainfall probability density functions (PDFs) and rainfall spatial-temporal distributions, likely because WRF captured the TWI properties well. The historical simulation outputs are available at hourly resolution for near surface (2-dimensional) fields and for the 3-dimensional atmosphere.

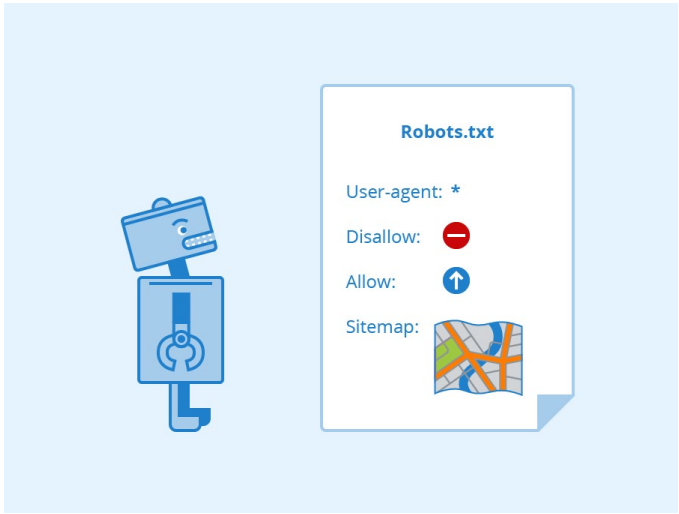
DOI:
https://doi.org/10.5065/7c2d-bg23


```
{
  "@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 hydrological cycle over th
  "url" : "https://doi.org/10.5065/7c2d-bg23",
  "sameAs" : "https://doi.org/10.5065/7c2d-bg23",
  "isAccessibleForFree" : "true"
}
```

Future Work for Sitemap/JSON -LD

Adding Robots.txt to Direct Search Engines to our sitemap.xml





BACKGROUND
Search Engine Faceting

Feature #3: Faceting

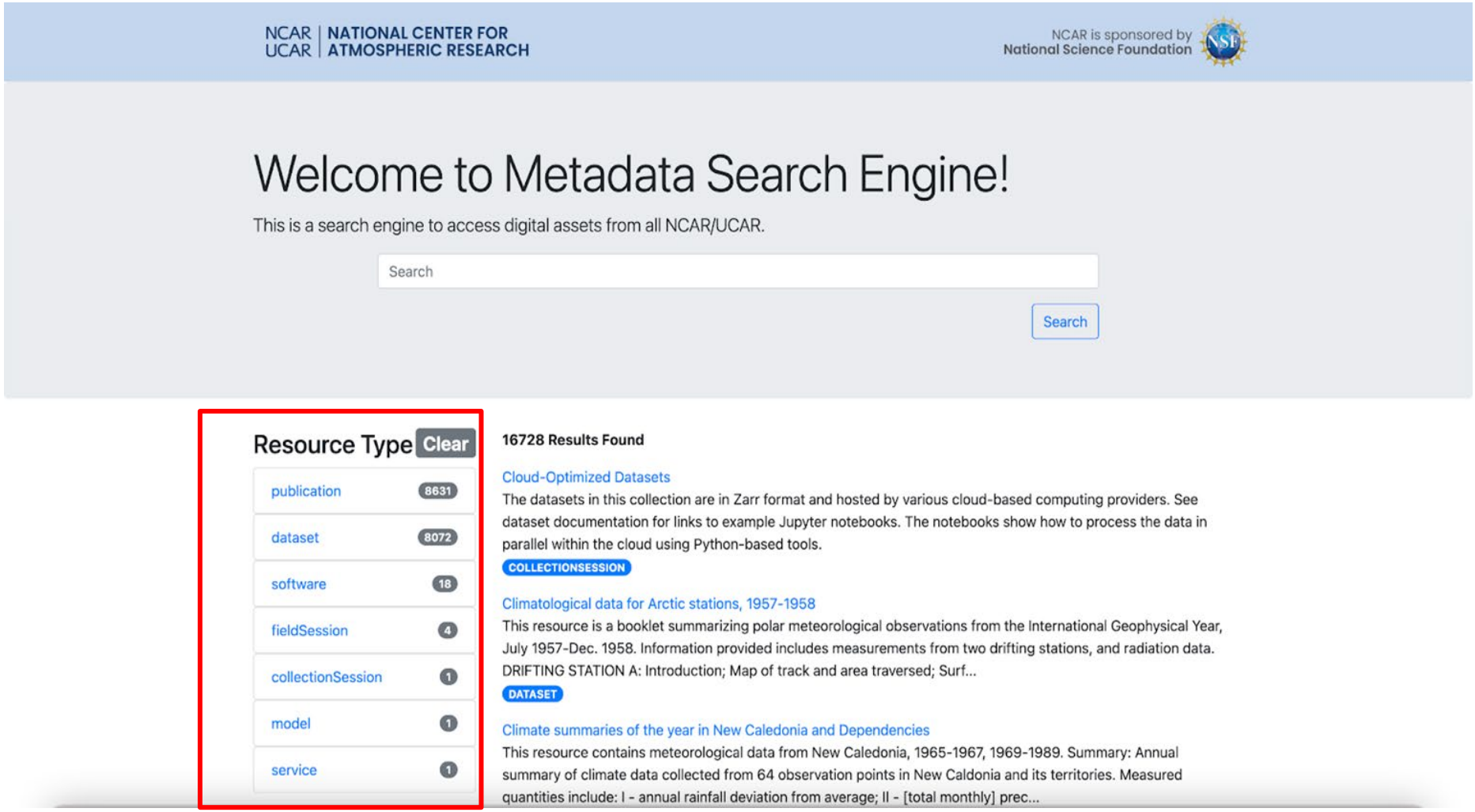
Why need?

- We want to be able to filter metadata search results by their resource type

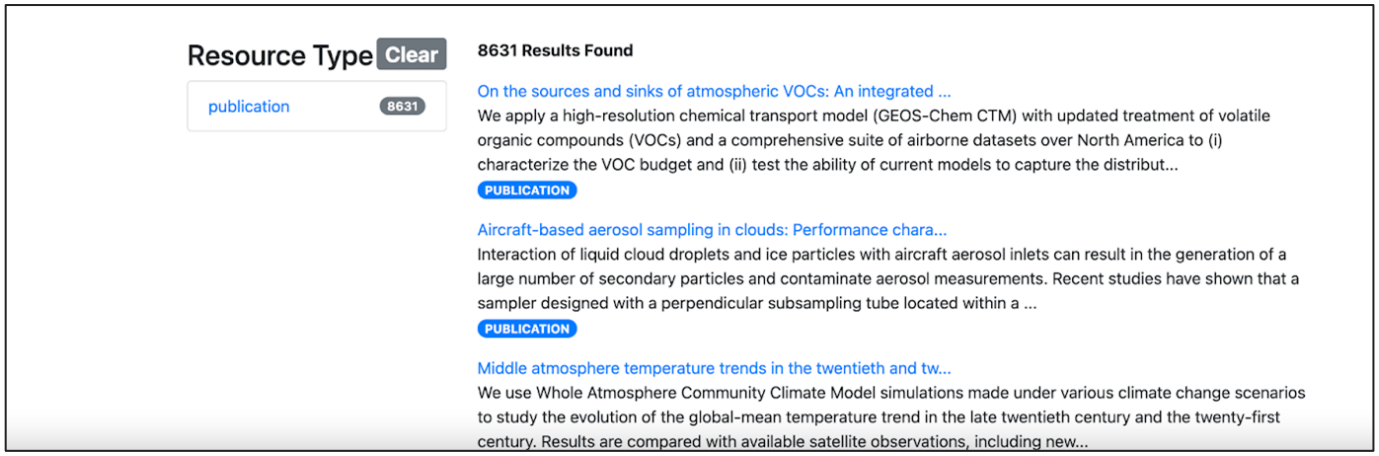

```
{
  "id":"ea0f3e13-8ec9-4424-bb8c-8fc4d79edce3",
  "title":"MetPy",
  "description":"MetPy is a collection of tools in Python for reading, visualizing, and performing calculations with weather data. The space MetPy a
  "doi":"https://www.unidata.ucar.edu/software/metpy",
  "keywords":["Software"],
  "resource_type":"software",
  "authors":["Ryan May"],
  "author_emails":["rmay@ucar.edu"],
  "authoritative_source_url":"https://www.unidata.ucar.edu/software/metpy",
  "authoritative_source_location_on_disk":"/Users/sduishebaeva/Java/xml/test-pull-method/test12.xml",
  "authoritative_source_md5":"d12d65916718529bc4e9e8ed1e9e7c28",
  "github_xml_url":"https://github.com/dsabira/test-pull-method.git/blob/main/test12.xml",
  "is_valid":true,
  "index_timestamp":"2022-07-15T17:11:34.267Z",
  "_version_":1738439492126113792},
{
  "id":"edu.ucar.rda::ds010.0",
  "title":"Daily Northern Hemisphere Sea Level Pressure Grids, continuing from 1899",
  "description":"The 5-degree latitude/longitude grids contained in this dataset make up the longest continuous set of daily gridded Northern Hemisp
  "doi":"https://doi.org/10.5065/7NB6-RJ33",
  "keywords":["EARTH SCIENCE > ATMOSPHERE > ATMOSPHERIC PRESSURE > SEA LEVEL PRESSURE"],
  "resource_type":"publication",
  "authoritative_source_url":"https://doi.org/10.5065/7NB6-RJ33",
  "authoritative_source_location_on_disk":"/Users/sduishebaeva/Java/xml/test-pull-method/test7.xml",
  "authoritative_source_md5":"84df1f65ac71e181212602691f04f0d1",
  "github_xml_url":"https://github.com/dsabira/test-pull-method.git/blob/main/test7.xml",
  "is_valid":true,
  "index_timestamp":"2022-07-15T17:11:34.987Z",
  "_version_":1738439492881088512},
```

Results & Future Work
Search Engine Faceting

Faceting: Resource Type in the UI

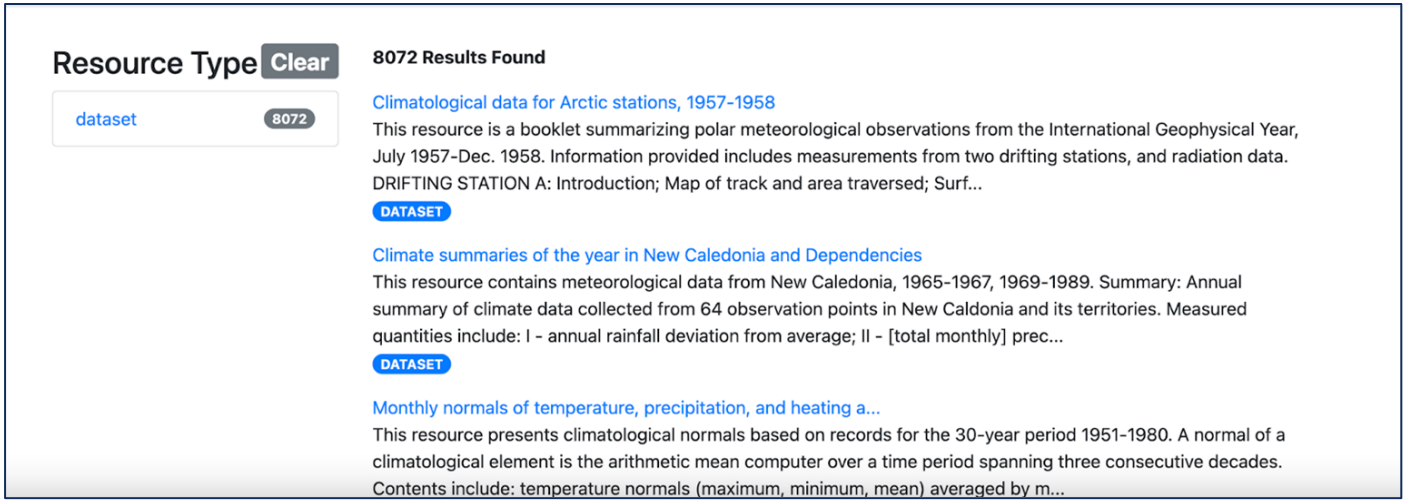


Faceting: It Works!



Future Work:

- More faceting options such as labs, keywords, authors etc.
- Multi-select option



ACKNOWLEDGEMENTS

Thank You

NCAR and CISL

SIParCS Mentors:

- Nathan Hook
- Saquib Aziz Khan
- Eric Nienhouse

SIParCS Program Leads:

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

Project Partner:

- Teagan Johnson

And everyone else involved with the SIParCS program this summer.

Resources

- https://thenounproject.com/icons/ - Most of the images/icons used
- https://www.searchenginejournal.com/upgrade-to-json-ld-structured-data/319327/ - Information about JSON-LD
- https://www.seobility.net/en/wiki/Robots.txt - Images/information about the Sitemap

QUESTIONS?