EOSDIS
NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

Google Dataset Search & CMR

WGISS-47

Doug Newman
EED-2 Data Use Architect
douglas.j.newman@nasa.gov

This work was supported by NASA/GSFC under Raytheon Co. contract number NNG15HZ39C.

Agenda

- **Concept**
- **Implementation**
- **Results**
- **Tooling**
- **Conclusions**
- **Future**

CONCEPT

What was at stake

Traditionally it has been difficult to get your collection landing pages near the top of a Google search.

First exposure

‘**Facilitating Dataset Discovery using new developments within schema.org**’ @ Summer ESIP 2013, Peter Fox

Adding schema.org ‘Dataset’ markup to your collection landing pages *might* solve the problem

IMPLEMENTATION

The players

CMR

The Common Metadata Repository (CMR) is a high-performance, high-quality, continuously evolving metadata system that catalogs Earth Science data and associated service metadata records. These metadata records are registered, modified, discovered, and accessed through programmatic interfaces leveraging standard protocols and APIs.

https://cmr.earthdata.nasa.gov/search/

Schema.org semantic markup

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

https://schema.org/

In collaboration

CMR collection landing pages, rendered in HTML can provide markup conforming to the schema.org standard to convey semantic information to commercial search engines and, in theory, allow users to discover and acquire NASA's earth science data in an intuitive and efficient manner.

Strategy

1. Comprehensive site map of cmr.earthdata.nasa.gov
2. RDFa schema.org tags embedded into each collection’s html landing page.

Schema.org mappings

CMR concept	Schema.org dataset concept
Entry Title	Name
Short Name	Alternate Name
Version ID	Version
Description	Description
Spatial extent of dataset	Spatial Coverage
Temporal extent of dataset	Temporal Coverage
GCMD Science Keywords for dataset	Keywords
DOI	DOI
Author/DAAC/POC	Citation
Creation date	Date Created
Last update date	Date Modified
Provider	Provider
Online Access URLs	URL of type distribution
Other URLs	URL

Example markup (1 of 3)

```
<div itemscope itemtype="http://schema.org/Dataset">
  <meta itemprop="name" content ="MODIS/Terra Calibrated Radiances 5-Min L1B Swath 250m V006"/>
  <meta itemprop="alternateName" content ="MOD02QKM_6"/>
  <meta itemprop="version" content="6"/>
  ...
</div>

<span itemprop='description'>The MODIS/Terra Calibrated Radiances 5-Min L1B Swath 250m...</span>

<meta itemprop="spatialCoverage">
  <div vocab="http://schema.org/" typeof="Place">
    <div property="geo" typeof="GeoShape">
      <meta property="box" content="90.0 -180.0 90.0 180.0" />
    </div>
  </div>
</meta>

<time itemprop="temporalCoverage" datetime="2000-02-24T00:00:00.000Z/..">2000-02-24 to present</time>

<meta itemprop="keywords" content="EARTH SCIENCE,SPECTRAL/ENGINEERING,INFRARED WAVELENGTHS"/>
```

EOSDIS

Example markup (2 of 3)

```
<li itemprop="identifier" itemscope="" itemtype="http://schema.org/PropertyValue">
  <h5>DOI</h5>
  <p>
    <meta itemprop="propertyID" content="DOI">
    <span itemprop="value">10.5067/MODIS/MOD02QKM.006</span>
  </p>
</li>

<li itemprop="citation" itemtype="http://schema.org/CreativeWork" itemscope>
  <h5 itemprop="headline">MODIS/Terra Calibrated Radiances 5-Min L1B Swath 250m V006</h5>
  <span itemprop="author">MCST Team</span>
  <span itemprop="publisher">L1 and Atmosphere Archive and Distribution System (LAADS)</span>
  <a itemprop="url" href="http://example.com">https://dx.doi.org/10.5067/MODIS/MOD02QKM.006</a>
</li>

<time itemprop='dateCreated' datetime='2012-11-05T00:00:00.000Z'>2012-11-05T00:00:00.000Z</time>
<time itemprop='dateModified' datetime='2017-12-28T00:00:00.000Z'>2017-12-28T00:00:00.000Z</time>

<h5 itemprop='provider'>GSFC</h5>
```

Example markup (3 of 3)

```
<a href="https://modaps.nascom.nasa.gov/services/about/product_descriptions_terra.html" itemprop="url">
  https://modaps.nascom.nasa.gov/services/about/product_descriptions_terra.html
</a>
<a href="https://ladsweb.modaps.eosdis.nasa.gov" itemprop="distribution" itemscope="itemscope" itemtype="http://schema.org/DataDownload">
  <meta itemprop="contentUrl" content="https://ladsweb.modaps.eosdis.nasa.gov"/>
  https://ladsweb.modaps.eosdis.nasa.gov
</a>
```

EOSDIS

RESULTS

The long wait

We waited 5 years.

We continued to report back to ESIP and made modifications based on the evolution of schema.org

But no tangible results from searches of Google.

Finally…

In 2018, NASA, along with other agencies, were asked to help with the development of a bespoke Google search engine for science data.

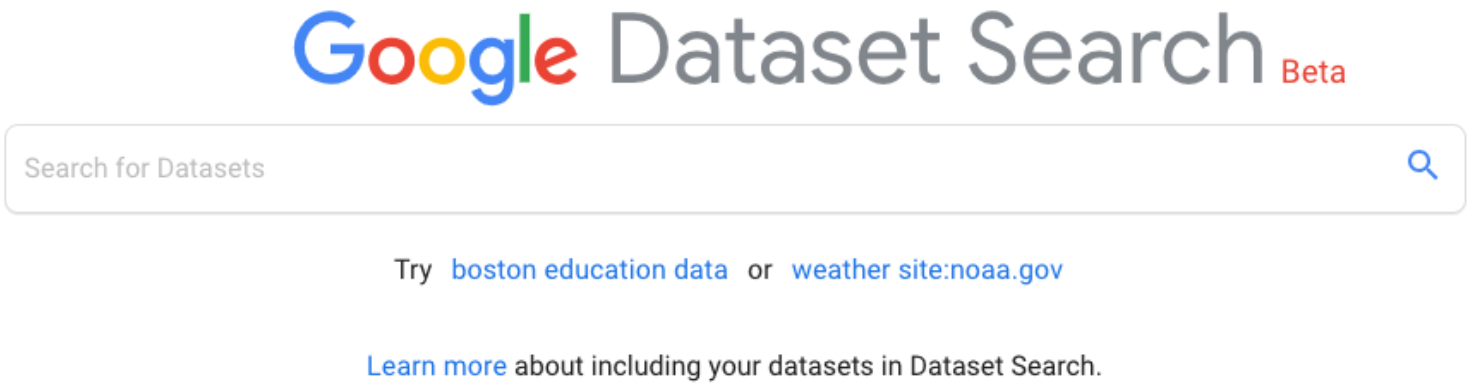
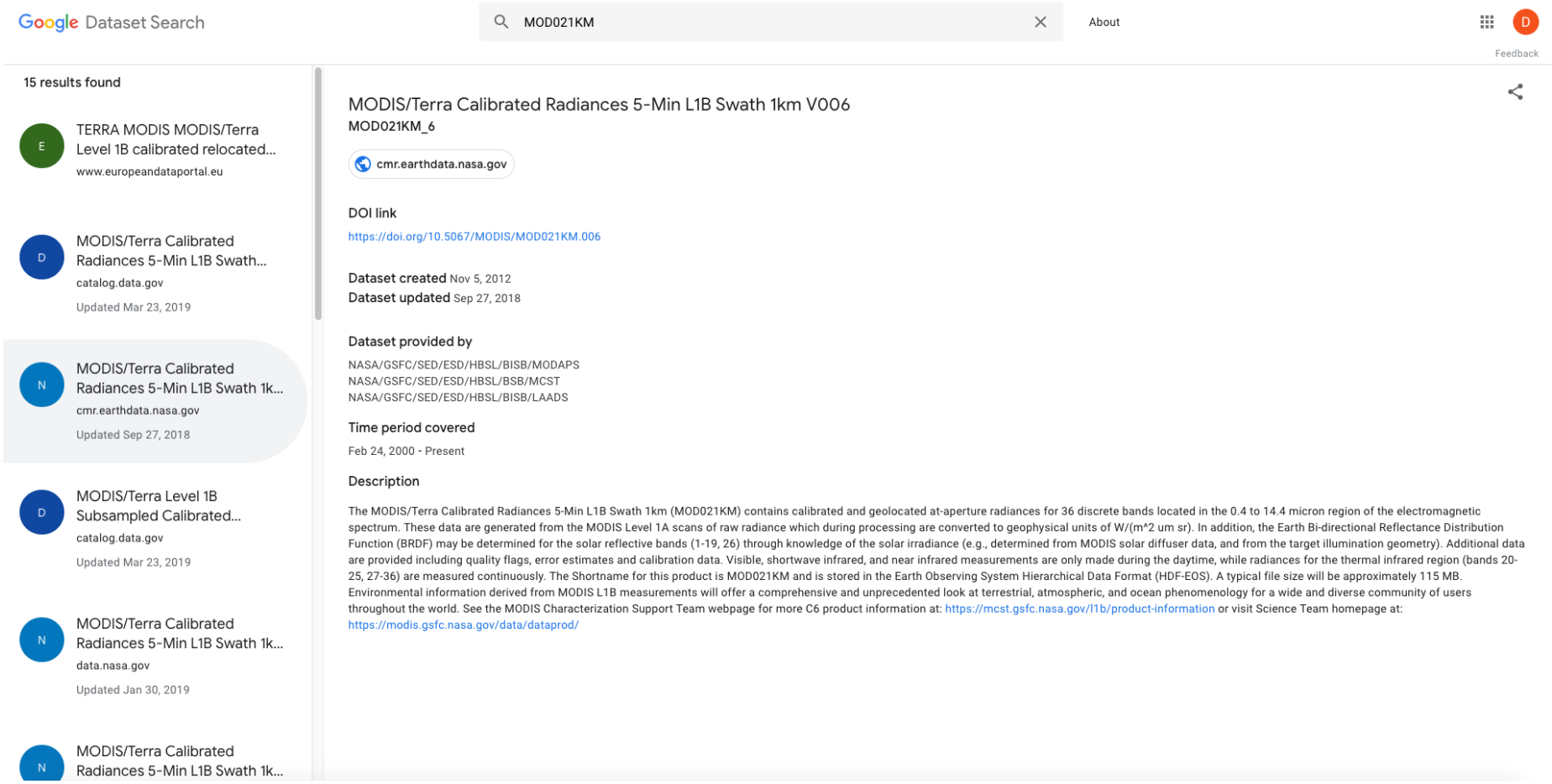


Image source: https://toolbox.google.com/datasetsearch

EOSDIS

An example from CMR



EOSDIS

Testing our markup

Schema org element	Search query	Comments
Alternate Name	MOD021KM	
Name	MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km	
Version	MOD021KM 6.1	
Temporal	MOD021KM 1999	No results
Temporal	MOD021KM 2018	Results
DOI	10.5067/MODIS/MOD021KM.061S	Single result
Science Keywords	VISIBLE RADIANCE MOD021KM	Results
Science Keywords	ultraviolet radiance MOD021KM	No results

EOSDIS

TOOLING

Google Dataset Guidelines

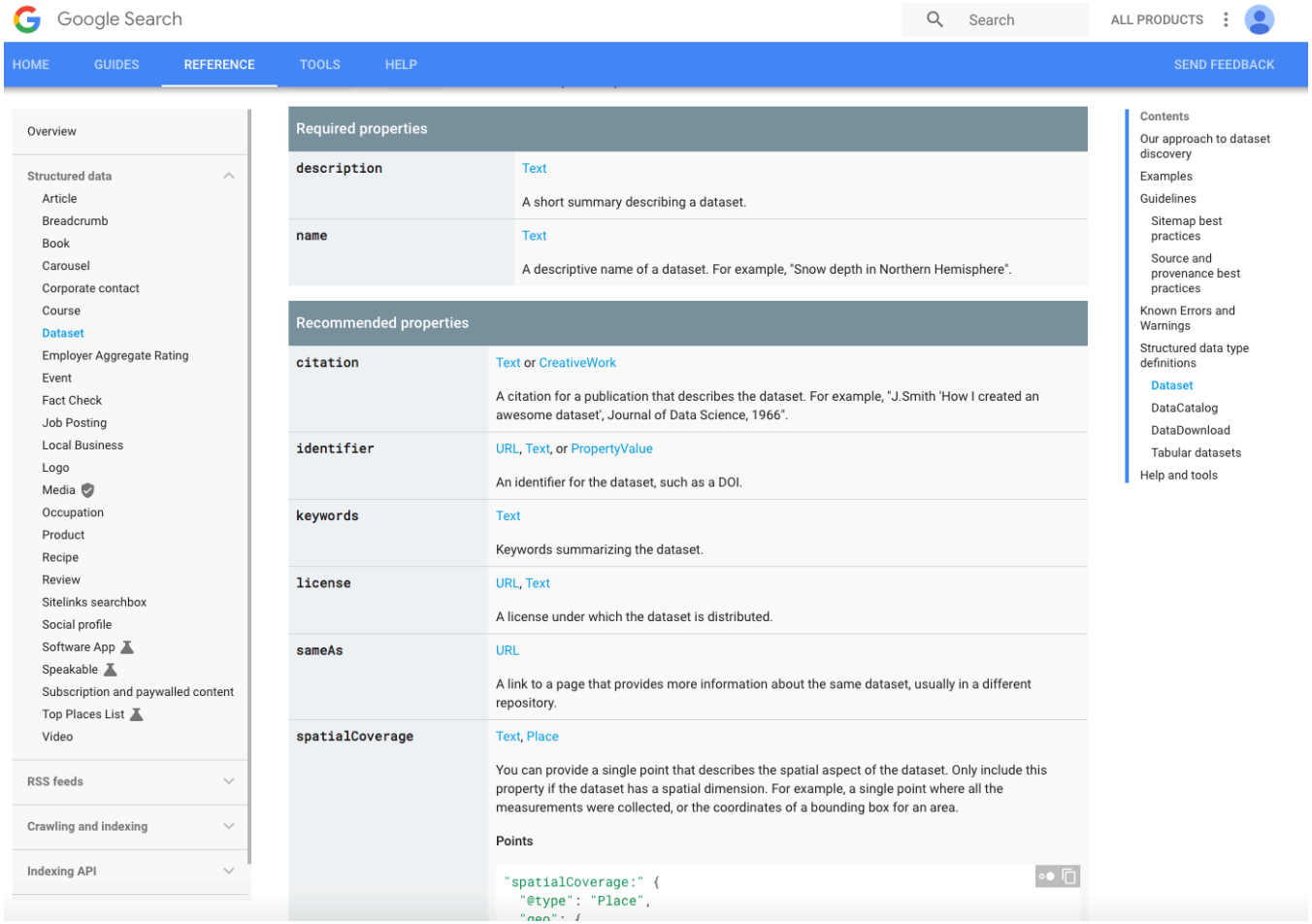


Image source: https://developers.google.com/search/docs/data-types/dataset)

EOSDIS

Structured Data Testing Tool

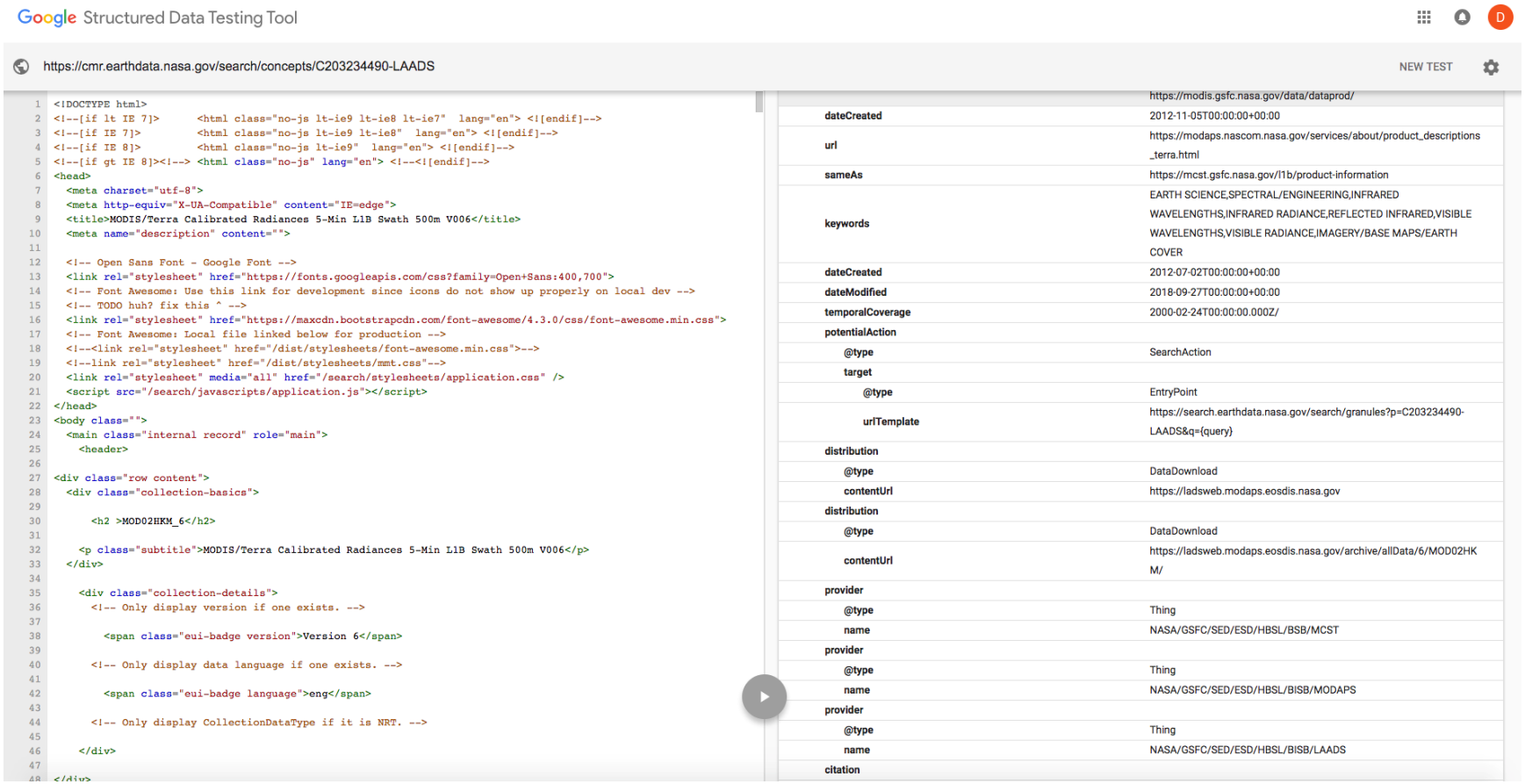


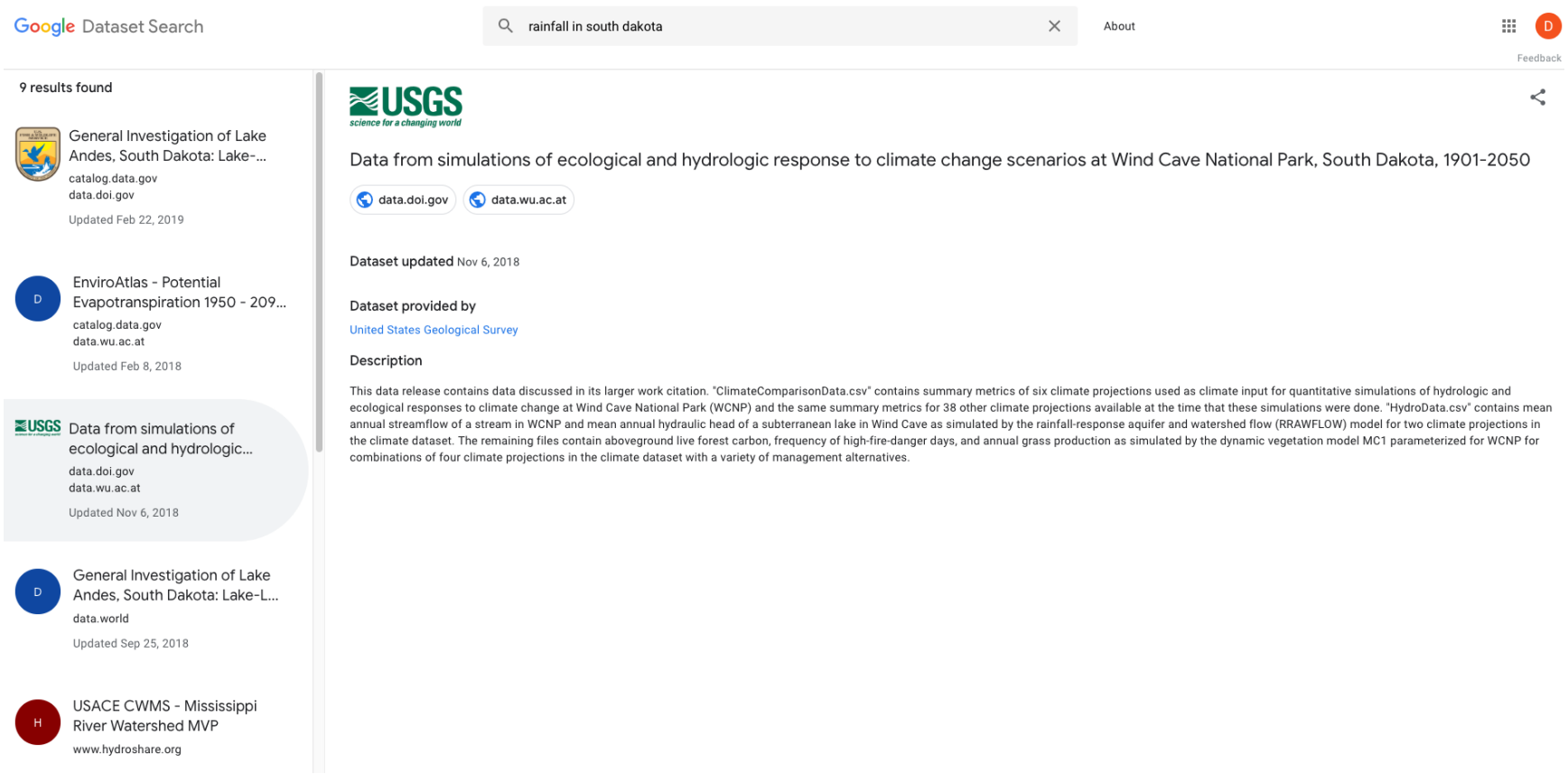
Image source: https://search.google.com/structured-data/testing-tool

EOSDIS

CONCLUSIONS

Google’s Dataset Search Tool provides an excellent entry point for earth data science user data discovery.

Rainfall example



EOSDIS

FUTURE

Upstream content providers

CMR provides metadata to both nasa.data.gov and data.gov

Google also ranks their search results higher than that of CMR.

CMR have begun an effort to increase the semantic content of nasa.data.gov and, by extension, data.gov to resolve this problem.

Search to Dataset Search

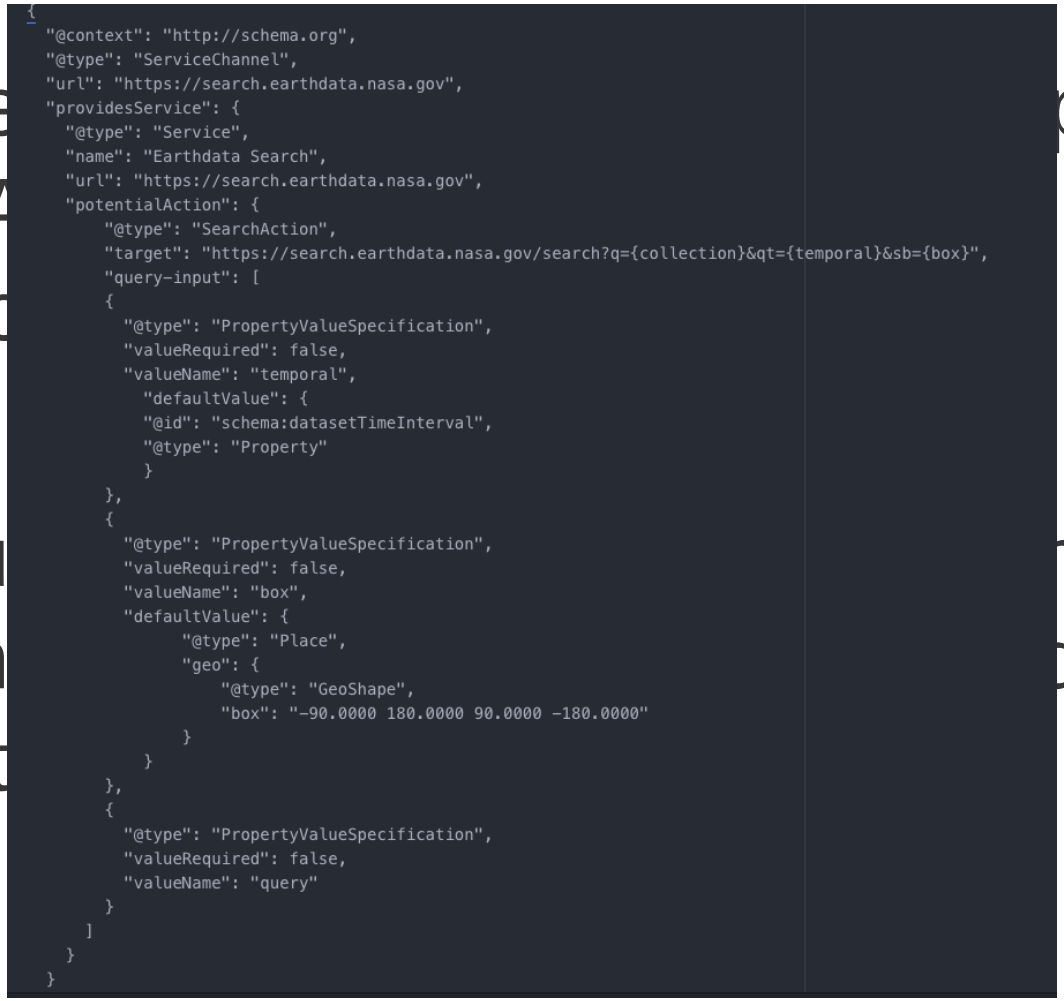
There is no concrete link between Google Search and Google Dataset Search.

It should emulate Google Jobs, for example.

Search Action usage

When se provided with an A Amtrak.c

One cou nship between d Earthdat



```
{
  "@context": "http://schema.org",
  "@type": "ServiceChannel",
  "url": "https://search.earthdata.nasa.gov",
  "providesService": {
    "@type": "Service",
    "name": "Earthdata Search",
    "url": "https://search.earthdata.nasa.gov",
    "potentialAction": {
        "@type": "SearchAction",
        "target": "https://search.earthdata.nasa.gov/search?q={collection}&qt={temporal}&sb={box}",
        "query-input": [
        {
          "@type": "PropertyValueSpecification",
          "valueRequired": false,
          "valueName": "temporal",
            "defaultValue": {
            "@id": "schema:datasetTimeInterval",
            "@type": "Property"
            }
        },
        {
          "@type": "PropertyValueSpecification",
          "valueRequired": false,
          "valueName": "box",
          "defaultValue": {
                "@type": "Place",
                "geo": {
                    "@type": "GeoShape",
                    "box": "-90.0000 180.0000 90.0000 -180.0000"
                }
            }
        },
        {
          "@type": "PropertyValueSpecification",
          "valueRequired": false,
          "valueName": "query"
        }
      ]
    }
  }
```

QUESTIONS

This work was supported by NASA/GSFC under Raytheon Co. contract number NNG15HZ39C.

Raytheon

in partnership with