

Schema.org: Improving access to data through a standardized language

Doug Fils, Consortium for Ocean Leadership

Adam Shepherd, Biological and Chemical Oceanography Data Management Office

Bryce Mecum, DataONE & NCEAS/UCSB

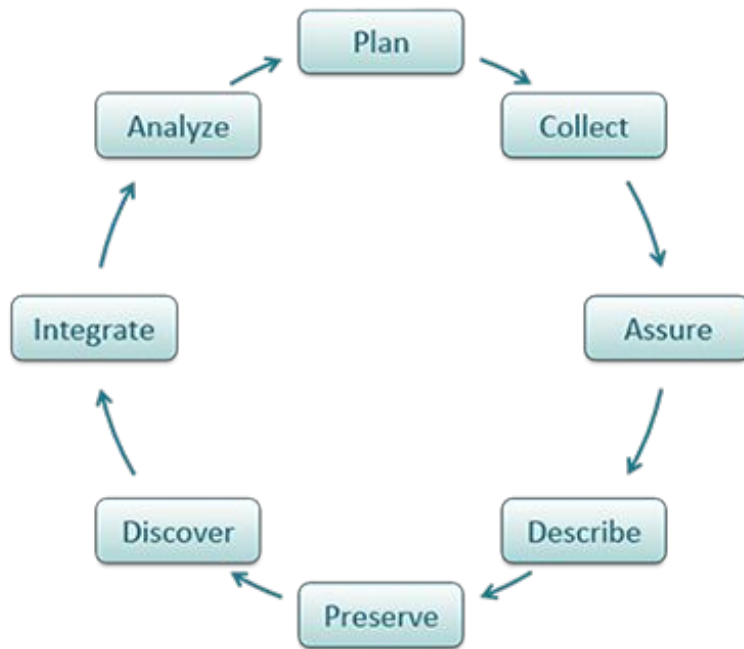


Problem statement

Finding data on the web is hard

It's getting better

It could be easier/better



How do we find data as researchers?

- Colleagues
- re3data (<https://www.re3data.org>)
- DataONE Search (<https://search.dataone.org>)
- DataCite Search (<https://search.datacite.org>)
- Other?
- ...*Google?*

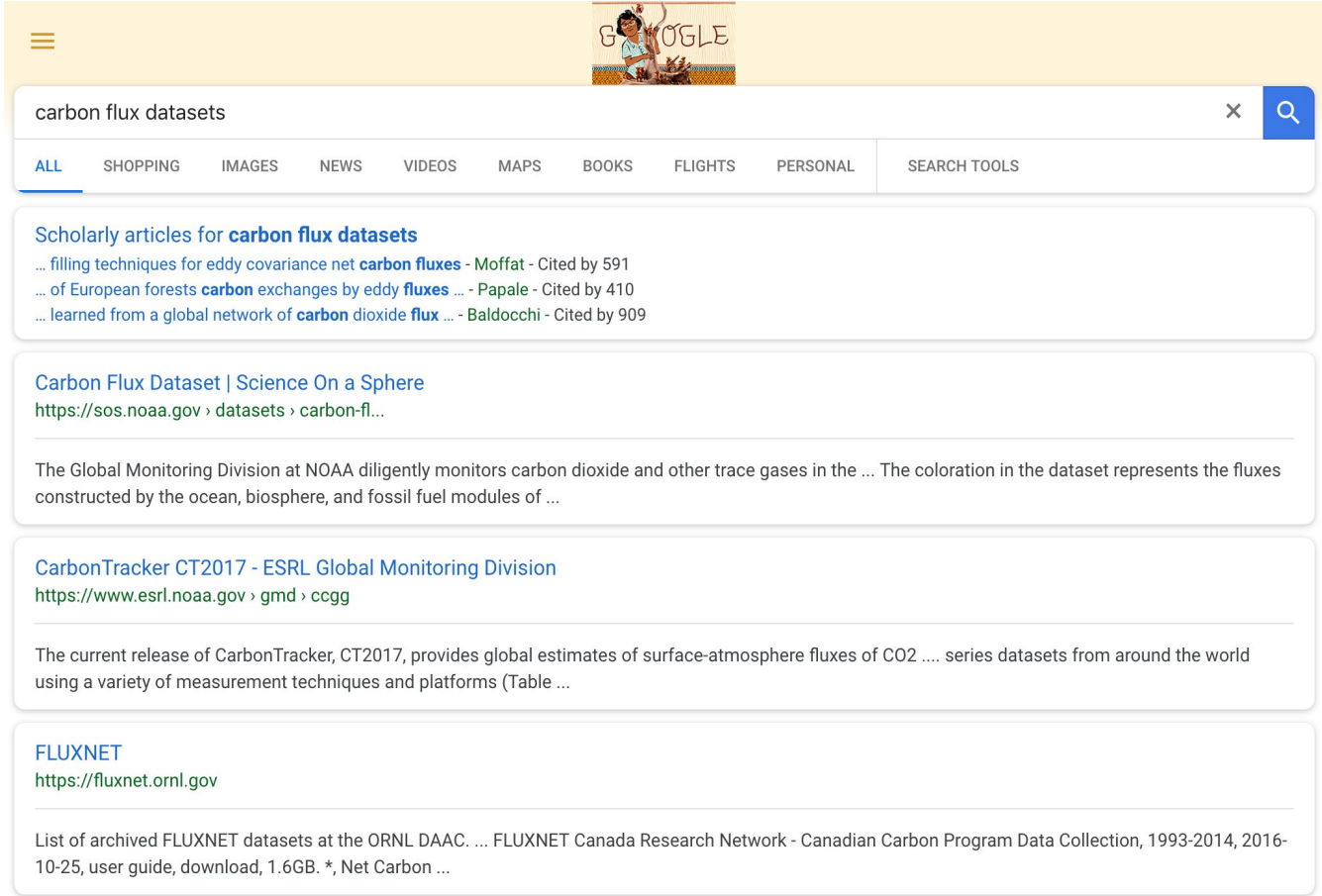
Philosophy time

If it isn't on Google, does it exist?

Philosophy time

~~If it isn't on Google, does it exist?~~

If **a dataset** isn't on Google, does it exist?



The image shows a Google search results page for the query "carbon flux datasets". At the top, there is a yellow header with the Google logo and a search bar containing the text "carbon flux datasets". Below the search bar, there are navigation tabs for "ALL", "SHOPPING", "IMAGES", "NEWS", "VIDEOS", "MAPS", "BOOKS", "FLIGHTS", "PERSONAL", and "SEARCH TOOLS". The "ALL" tab is selected. The search results are displayed in a list of cards. The first card is titled "Scholarly articles for carbon flux datasets" and lists three articles with their authors and citation counts. The second card is titled "Carbon Flux Dataset | Science On a Sphere" and includes a URL and a brief description of the dataset. The third card is titled "CarbonTracker CT2017 - ESRL Global Monitoring Division" and includes a URL and a brief description of the dataset. The fourth card is titled "FLUXNET" and includes a URL and a brief description of the dataset.

carbon flux datasets

ALL SHOPPING IMAGES NEWS VIDEOS MAPS BOOKS FLIGHTS PERSONAL SEARCH TOOLS

Scholarly articles for **carbon flux datasets**

- ... filling techniques for eddy covariance net **carbon fluxes** - Moffat - Cited by 591
- ... of European forests **carbon** exchanges by eddy **fluxes** ... - Papale - Cited by 410
- ... learned from a global network of **carbon** dioxide **flux** ... - Baldocchi - Cited by 909

Carbon Flux Dataset | Science On a Sphere
<https://sos.noaa.gov/datasets/carbon-fl...>

The Global Monitoring Division at NOAA diligently monitors carbon dioxide and other trace gases in the ... The coloration in the dataset represents the fluxes constructed by the ocean, biosphere, and fossil fuel modules of ...

CarbonTracker CT2017 - ESRL Global Monitoring Division
<https://www.esrl.noaa.gov/gmd/ccgg>

The current release of CarbonTracker, CT2017, provides global estimates of surface-atmosphere fluxes of CO2 series datasets from around the world using a variety of measurement techniques and platforms (Table ...

FLUXNET
<https://fluxnet.ornl.gov>

List of archived FLUXNET datasets at the ORNL DAAC. ... FLUXNET Canada Research Network - Canadian Carbon Program Data Collection, 1993-2014, 2016-10-25, user guide, download, 1.6GB. *, Net Carbon ...

iphone

All News Shopping Images Videos More

About 3,630,000,000 results (0.52 seconds)

Shop iPhone



Apple iPhone XR -
64GB - Blue - ...

\$25.00/mo
For 30 months
AT&T



Apple iPhone XR -
128GB - ...

\$26.67/mo
For 30 months
AT&T



Apple iPhone XS -
256GB - Gold - ...

\$38.34/mo
For 30 months
AT&T



Apple iPhone XS
Max - 256GB - ...

\$41.67/mo
For 30 months
AT&T



Apple iPhone 8
Plus - 64GB - Go...

\$23.34/mo
For 30 months
AT&T



Google does have
structured knowledge
about *some* things...

carbon flux

All News Shopping Images Videos More

About 3,630,000,000 results (0.52 seconds)

Find Carbon Flux data



Net Ecosystem Carbon Flux

From: United States
Geologic Survey
Updated: Jun 8, 2018
Format: GeoTIFF



Goddard
Space Flight Center

Carbon Monitoring System Flux from the Net Ecosy...

From: NASA Goddard
Space Flight Center
Updated: Aug 1, 2018
Format: NetCDF v4



Hydrological and Energy Surface and Atmospheri...

From: LTER
Updated: October 12, 2018
Format: CSV

DataCite
FIND, ACCESS, AND REUSE DATA

Estimating carbon fluxes using satellite data ...

From: DataCite
Published: 2017
Format: NetCDF



...but what if it had structured knowledge about **data**?



<https://toolbox.google.com/datasetsearch>

Google Dataset Search Beta

Search for Datasets



Try [boston education data](#) or [weather site:noaa.gov](#)



Monthly Weather Review

data.nodc.noaa.gov
catalog.data.gov

Updated May 2, 2013



Mariners Weather Log

data.nodc.noaa.gov
catalog.data.gov
+1more

Published 1957



Daily Weather Records

data.nodc.noaa.gov
catalog.data.gov
+1more

Published Dec 1, 2013



Oil Rig Weather Observations

data.nodc.noaa.gov
catalog.data.gov
+1more

Updated May 2, 2013



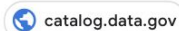
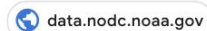
Surface Weather Observations

data.nodc.noaa.gov



Monthly Weather Review

gov.noaa.ncdc:C01044



Dataset created Mar 15, 2011

Dataset updated May 2, 2013

Dataset published Mar 15, 2011

Dataset provided by

[National Oceanic and Atmospheric Administration](#)

Time period covered 1914 - 1949

Area covered

United States of America, Pacific Ocean, North Pacific Ocean

Description

Supplements to the Monthly Weather Review publication. The Weather Bureau published the Monthly weather review Supplement irregularly from 1914 to 1949. The Supplement replaced numerous independent series of bulletins that the Bureau published before 1914. The Supplements featured contributions to the science of meteorology and weather forecasting that were too voluminous to publish in the regular Monthly weather review. The Bureau never published no. 43. The Monthly Weather Review series has also been scanned, and is hosted by the American Meteorological Society, which assumed publication in 1974.



How does this work?

For independent repositories:

1. Take your existing dataset landing pages
2. Add Schema.org markup (JSON-LD) into your <head> tags for each page
3. Wait for Google to re-crawl your pages

For DataONE member nodes:

1. We're looking into doing this across the federation on <https://search.dataone.org>

How does this work?

- Many repositories are already doing this
- If you're a researcher, you may just need to get in touch with your data manager or repository operator to ask about this

```
{
  "@context": {
    "@vocab": "http://schema.org/"
  },
  "@type": "Dataset",
  "name": "Dynamic energy budget model parameter estimation files for ...",
  "description": "Matlab files for use with the DEBtool package for ...",
  "variableMeasured": [
    {
      "@type": ["PropertyValue"],
      "name": "length",
      "description": "Snout vent length (mm)"
    },
    {
      "@type": ["PropertyValue"],
      "name": "weight",
      "description": "total wet weight (g)"
    }
  ]
}
```

This is just a sample of a few properties we can describe

```

<!doctype html>
<html lang="en">
  <head>
    <title>Dynamic energy budget model parameter estimation files for ... </title>
    <script type="application/ld+json">
      {
        "@context": {
          "@vocab": "http://schema.org/"
        },
        "@type": "Dataset",
        "name": "Dynamic energy budget model parameter estimation files for ...",
        "description": "Matlab files for use with the DEBtool package for ...",
        "variableMeasured": [
          {
            "@type": ["PropertyValue"],
            "name": "length",
            "description": "Snout vent length (mm)"
          },
          {
            "@type": ["PropertyValue"],
            "name": "weight",
            "description": "total wet weight (g)"
          }
        ]
      }
    </script>
  </head>
  <body>
    ...
  </body>
</html>

```





This is just a sample of a few properties we can describe

Schema.org Summary

- Very easy to adopt (insert JSON-LD into HTML pages you already have)
- Lightweight but extensible vocabulary (Schema.org)
 - Easy to convert from richer XML schemas (ISO19139, EML, etc) we already have
- We can teach Google about data
- Important outside of Google Dataset Search: Schema.org could be a lingua franca for structured knowledge about web content
- *Still experimental*. Google isn't using most of the Schema.org markup **yet**.

Project 418: Goals

Worked with NSF data facilities to leverage schema.org for dataset *description, indexing* and *discovery*

GOAL	Describing 	Publishing 	Indexing 	Serving 
STATUS	<p>P418 Vocabulary approaches developed, now working with ESIP on governance and evolution</p>	<p>Worked with facilities to adapt approach to their existing metadata workflow and software.</p> <p>10 NSF facilities publishing *</p>	<p>Code developed to collect and index the descriptions. Indexes include: text, spatial and graph.</p>	<p>Geodex.org, example notebooks and APIs.</p>

- Code at: <https://github.com/earthcubearchitecture-project418>
- Implementation at: <https://geodex.org/>
- * Several now part of Google Dataset Search
- Done in collaboration with a larger community working on these approaches

Publishing:

Done in collaboration with the facilities. Focused mostly on schema.org type Dataset with an eye toward future extension work. Vocabulary work now at ESIP

Indexing (summoning)

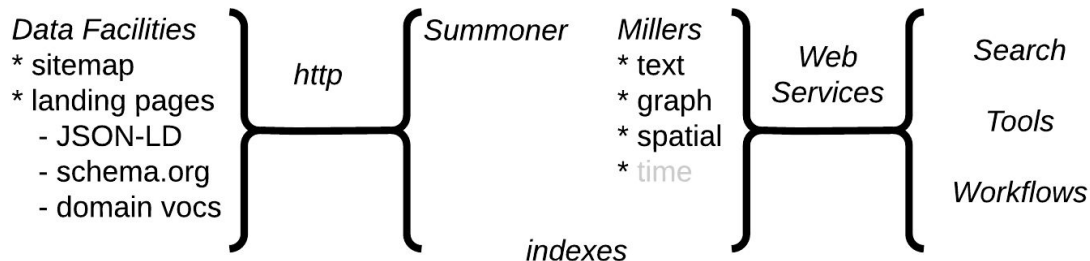
Go based code “gleaner (summoner)” built that pulls the JSON-LD based schema.org from resources. Driven by sitemap files.

Indexing (milling)

Go based code “gleaner (miller)” is a set of patterns for adding different indexing (“milling”) workflows to work on the summoned code. Main ones were spatial, text and graph. Also have SHACL, alternative indexing and other miller options in the works.

Serving

APIs and sample interface at <https://geodex.org> using indexes from millers.

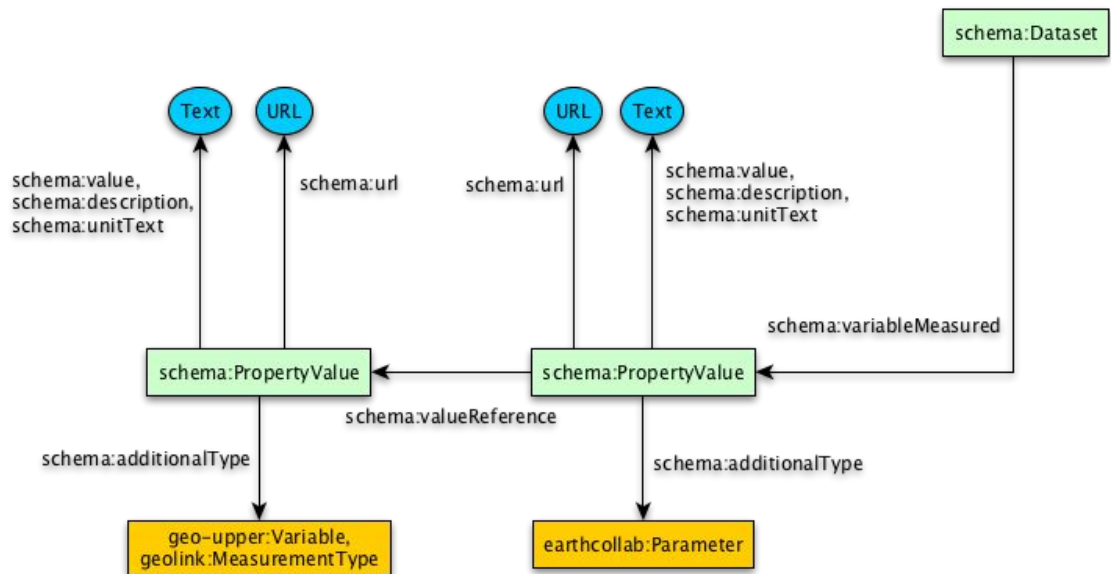




<https://github.com/earthcubearchitecture-project418/p418Vocabulary>

Using schema.org as a basis with a focus on type Dataset. Then providing example and reference implementation of using external vocabularies to address domain specific needs.

1. To produce quality schema.org markup with additional extensions to schema.org classes to help improve harvesting technologies.
2. Produced markup will pass the [Google Structured Data Testing Tool](#) with 0 errors.

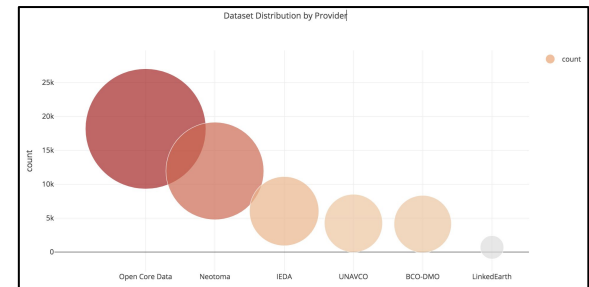
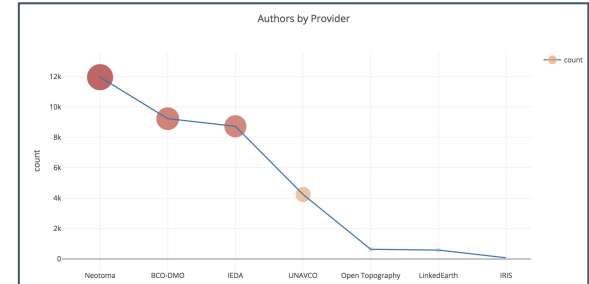


Vocabulary Use - Google Recommended

Dataset Properties	Google Requires / Recommends	Provider Usage	Dataset Coverage	
			Implemented	Overall
@context	Required. Set @context to "http://schema.org/"	80%	omitted ending slash: 'http://schema.org'	
@type	Required. Set @type to "Dataset"	100%	47,650 datasets	n/a
name	Required. A descriptive name	80%	99.9%	73%
description	Required. A short summary	70%	97%	69%
url	Recommended.	70%	100%	62%
citation	Recommended.	60%	100%	36%
keywords	Recommended.	70%	99.9%	66%
spatialCoverage	Recommended.	80%	92%	91%
temporalCoverage	Recommended.	10%	15%	<1%
variableMeasured	Recommended.	30%	83%	40%
version	Recommended.	40%	95%	25%
sameAs	Recommended. Same data, different URL.	10%	100%	<1%

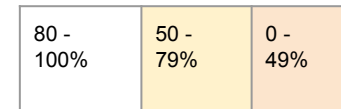
Vocabulary Use - P418 Recommended

Dataset Properties	Provider Usage		Dataset Coverage	
			Implemented	Overall
identifier	30%	10,556 datasets	100%	22%
author/creator/contributor	80%	28,765 datasets	98%	69%
funder (not awards)	30%	4,069 datasets	78%	9%
distribution	60%	45,221 datasets	100%	95%
license	70%	42,523 datasets	98%	89%
hasPart <i>ex: linking PhysicalSamples to Datasets</i>	10%	122 datasets	2%	<1%



"What about Data APIs?"

- **3 providers:** Search endpoints, SWAGGER, SPARQL, VoID, OGC CSW



Vocabulary Use - External Vocabularies

- Some providers used external vocabularies
 - EarthCube Building Blocks - EarthCollab & GeoLink
 - Datacite Ontology - DOIs and ORCID
 - ViVO Ontology - Datasets

Opportunity to improve search precision

- Geoscience Standard Names,
- SWEET Ontologies,
- GCMD Keywords,
- etc.



Future

Project 418 was a rapid “proof of concept”

Vocabulary work moving to ESIP

An EarthCube follow on project has been awarded:

- Code improvements to support better maintenance
- Make indexer easier for others to use
- Improve scaling
- Explore extension model of schema.org in greater detail

XSEDE

Extreme Science and Engineering
Discovery Environment

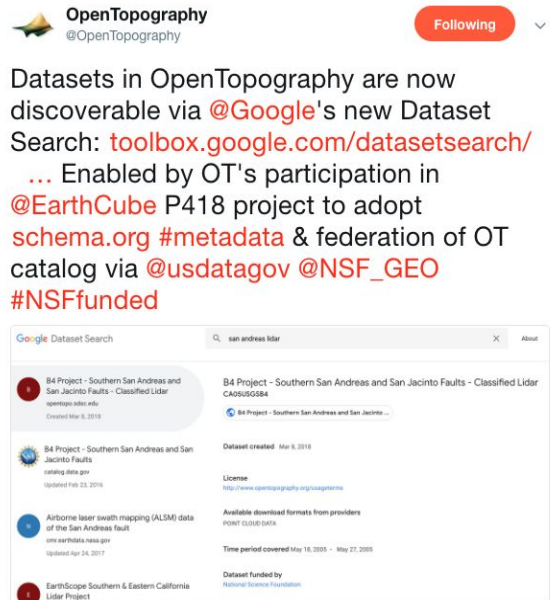
Special Thanks to XSEDE for hosting
and technical assistance.



BCO-DMO @BCODMO Following

Thanks to @EarthCube P418 project-Check out the @BCODMO data catalog now searchable through @google data search toolbox.google.com/datasetsearch/ ...

12:38 PM - 5 Sep 2018



OpenTopography @OpenTopography Following

Datasets in OpenTopography are now discoverable via @Google's new Dataset Search: toolbox.google.com/datasetsearch/ ... Enabled by OT's participation in @EarthCube P418 project to adopt schema.org #metadata & federation of OT catalog via @usdatagov @NSF_GEO #NSFfunded

3:37 PM - 5 Sep 2018

3:37 PM - 5 Sep 2018



1,113,210
7,087,380
47,650
54,665
599,960
~ 35k
~560k

Entities
Triples
Dataset
DataDownload
PropertyValue
Identifier
Dataset Variables