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
Schema.org: Improving access to data through a standardized language.
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?
Scholarly articles for carbon flux datasets

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 ...
Google does have **structured** knowledge about some things...
...but what if it had structured knowledge about data?
Schema.org: Improving access to data through a standardized language.

https://toolbox.google.com/datasetsearch
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
This is just a sample of a few properties we can describe
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
**Project 418: Goals**

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

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

- Code at: [https://github.com/earthcubearchitecture-project418](https://github.com/earthcubearchitecture-project418)
- Implementation at: [https://geodex.org/](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, test 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.

Data Facilities
- sitemap
- landing pages
  - JSON-LD
  - schema.org
  - domain vocs

Summoner

Millers
- text
- graph
- spatial
- time

Web Services

Search

Tools

Workflows
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

<table>
<thead>
<tr>
<th>Dataset Properties</th>
<th>Google Requires / Recommends</th>
<th>Provider Usage</th>
<th>Dataset Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Implemented</td>
<td>Overall</td>
</tr>
<tr>
<td>@context</td>
<td>Required. Set @context to &quot;<a href="http://schema.org/">http://schema.org/</a>&quot;</td>
<td>80%</td>
<td>n/a</td>
</tr>
<tr>
<td>@type</td>
<td>Required. Set @type to &quot;Dataset&quot;</td>
<td>100%</td>
<td>47,650 datasets</td>
</tr>
<tr>
<td>name</td>
<td>Required. A descriptive name</td>
<td>80%</td>
<td>99.9%</td>
</tr>
<tr>
<td>description</td>
<td>Required. A short summary</td>
<td>70%</td>
<td>97%</td>
</tr>
<tr>
<td>url</td>
<td>Recommended.</td>
<td>70%</td>
<td>100%</td>
</tr>
<tr>
<td>citation</td>
<td>Recommended.</td>
<td>60%</td>
<td>100%</td>
</tr>
<tr>
<td>keywords</td>
<td>Recommended.</td>
<td>70%</td>
<td>99.9%</td>
</tr>
<tr>
<td>spatialCoverage</td>
<td>Recommended.</td>
<td>80%</td>
<td>92%</td>
</tr>
<tr>
<td>temporalCoverage</td>
<td>Recommended.</td>
<td>10%</td>
<td>15%</td>
</tr>
<tr>
<td>variableMeasured</td>
<td>Recommended.</td>
<td>30%</td>
<td>83%</td>
</tr>
<tr>
<td>version</td>
<td>Recommended.</td>
<td>40%</td>
<td>95%</td>
</tr>
<tr>
<td>sameAs</td>
<td>Recommended. Same data, different URL.</td>
<td>10%</td>
<td>100%</td>
</tr>
</tbody>
</table>

https://developers.google.com/search/docs/data-types/dataset
Vocabulary Use - P418 Recommended

<table>
<thead>
<tr>
<th>Dataset Properties</th>
<th>Provider Usage</th>
<th>Dataset Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Implemented</td>
</tr>
<tr>
<td>identifier</td>
<td>30%</td>
<td>10,556 datasets</td>
</tr>
<tr>
<td>author/creator/contributor</td>
<td>80%</td>
<td>28,765 datasets</td>
</tr>
<tr>
<td>funder (not awards)</td>
<td>30%</td>
<td>4,069 datasets</td>
</tr>
<tr>
<td>distribution</td>
<td>60%</td>
<td>45,221 datasets</td>
</tr>
<tr>
<td>license</td>
<td>70%</td>
<td>42,523 datasets</td>
</tr>
<tr>
<td>hasPart</td>
<td>10%</td>
<td>122 datasets</td>
</tr>
</tbody>
</table>

"What about Data APIs?"
- 3 providers: Search endpoints, SWAGGER, SPARQL, VoID, OGC CSW

github: [ESIPFed/science-on-schema.org](https://github.com/ESIPFed/science-on-schema.org)
github: [earthcubearchitecture-project418/p418Vocabulary](https://github.com/earthcubearchitecture-project418/p418Vocabulary)
• 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.

github: earthcubearchitecture-project418/assay-data
Project 418: Status & 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

Special Thanks to XSEDE for hosting and technical assistance.