ESIP Schema.org cluster

Using the Schema.org vocabulary for FAIR Earth Science Data

by Adam Shepherd and Douglas Fils

Making the Major Facilities Data Lifecycle FAIR to Provide AI-Ready Data

March 1, 2022
CI Compass Cyberinfrastructure for NSF Major Facilities Workshop
ESIP Schema.org cluster
Utilizing the Schema.org vocabulary for FAIR Earth Science Data

Adam Shepherd  
Technical Director  
BCO-DMO

Douglas Fils  
Data Manager  
Consortium of Ocean Leadership
Findable
Accessible
Interoperable
Reusable
What is Schema.org?

➢ Information embedded in HTML
➢ Classifies meaning
➢ Terms from https://schema.org
➢ Aligns with W3C Recommendation

Data on the Web Best Practices
w3.org/TR/DWBP
What is Schema.org?

➢ Information embedded in HTML
➢ Classifies the meaning of text, images, forms, etc.
➢ Terms from https://schema.org
➢ Aligns with W3C Recommendation Data on the Web Best Practices w3.org/TR/DWBP
Schema.org is…

➢ collaborative, community activity
➢ MISSION: to create, maintain, and promote schemas
➢ meant to be extended
➢ easy for publishers
   ○ Many dialects ➔ JSON-LD, RDFa, Microdata
   ○ Many embedding strategies ➔ inline HTML, <head>, HTTP Header

Evolution of Structured Data on the Web by Guha & Brickley, doi:10.1145/2857274.2857276
Test Immediately

Rich Results Test
The official Google tool for testing your structured data to see which Google rich results

search.google.com/test/rich-results

Schema Markup Validator
Validate all Schema.org-based structured data that's embedded in web pages, without

validator.schema.org
Benefits

- Publish once (for multiple harvesters)

  - Global
    - Google, Bing, Yahoo!
    - Google Dataset Search (GDSS)
  - Sciences
    - DataONE
  - NSF Geosciences
    - EarthCube GeoCODES
Science-on-Schema.org

Shared publishing patterns for describing research data on your web pages using schema.org

Tags:
Schema.org, Structured Data, RDF

Want to learn more?
science-on-schema.org
Shared Publishing Patterns

➢ **Reliable, consistent federation**

➢ **Automate Validation**

**Organization**
The sources for geodex come mostly from collaboration with the EarthCube Council of Data Facilities (CDF).

**Providers**
CDF members who express their resources via structured data on the web approaches can be indexed.

**Indexing**
Geodex uses the gleaner program (gleaner.io) to build the index and (GROW) as a server. See the about section for more.

**Validation Example**
Dataset: Photosymbiosis in planktonic foraminifera across the Palaeocene-Eocene Thermal Maximum.

<table>
<thead>
<tr>
<th>Violation</th>
<th>Reason</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Dataset must have an ID</td>
<td></td>
</tr>
<tr>
<td>Warning</td>
<td>It is recommended that a Dataset includes a sameAs URL</td>
<td>Path: <a href="http://schema.org/sameAs">http://schema.org/sameAs</a></td>
</tr>
<tr>
<td>Warning</td>
<td>It is recommended that a Dataset indicates accessibility for free or</td>
<td>Path: <a href="http://schema.org/isAccessibleForFree">http://schema.org/isAccessibleForFree</a></td>
</tr>
<tr>
<td></td>
<td>otherwise</td>
<td></td>
</tr>
<tr>
<td>Violation</td>
<td>Dataset must have a version as Literal or Number</td>
<td>Path: <a href="http://schema.org/version">http://schema.org/version</a></td>
</tr>
<tr>
<td>Violation</td>
<td>Dataset identifiers must be a URL, Text or PropertyValue</td>
<td>Path: <a href="http://schema.org/identifier">http://schema.org/identifier</a></td>
</tr>
</tbody>
</table>

5 errors, 10/23 tests applied.
WHY another Guideline?

- Flat descriptions
  - How are things connected?
- Limited examples
- Endless ways to publish

Q: How to do we share patterns of use so that no one is left behind?
Examples & Drawings

A point, or coordinate, would be defined in this way:

```json
{
  "@context": {
    "@vocab": "http://schema.org/",
    "datacite": "http://purl.org/spar/datacite/"
  },
  "@type": "Dataset",
  "name": "Removal of organic carbon by natural bacterioplankton",
  "spatialCoverage": {
    "@type": "Place",
    "geo": {
      "@type": "GeoCoordinates",
      "latitude": 39.3280,
      "longitude": 120.1633
    }
  }
}
```

All other shapes are defined using the `schema:GeoShape`:

```json
{
  "spatialCoverage": {
    "@type": "Place",
    "geo": {
      "@type": "GeoShape",
      "line": "39.3280,120.1633,40.445,123.7878"
    }
  }
}
```
Release Workflow

Use Github Issues

- **Use Git Flow** methodology
  - Master branch
  - Changes are made to 'develop' branch
  - Merged into master at the time of 'release'

- Github Milestones group Issues into official releases
  - Each Issue starts as a 'feature' branch
  - Pull Request into 'develop' branch
  - Reviewed by community

- Release day merges 'develop' into 'master'
Architectural Decision Records

github.com/.../blob/1.1.0/decisions

**Goal:** Crystalize decisions into digestible docs

4 sections:

- Status
- Decision
- Context
- Consequences

Link to a Github Issue with the full conversation

Use SPDX license vocabulary for URIs

Discussion: https://github.com/ESPRed/science-on-schema.org/issues/47

**Status**

Accepted

**Decision**

Use SPDX license URIs to unambiguously specify the license for data and metadata use.

**Context**

Link a Dataset to its license to document legal constraints by adding a `schema:license` property. The guide recommends providing a URL that unambiguously identifies a specific version of the license used, but for many licenses it is hard to determine what that URL should be. Thus, we recommend that the license URL be drawn from the SPDX license list, which provides a curated list of licenses and their properties that is well maintained. For each SPDX entry, SPDX provides a canonical URL for the license (e.g., `http://spdx.org/licenses/CC0-1.0`), a unique `licenseId` (e.g., `CC0-1.0`), and other metadata about the license. Here’s an example using the SPDX license URI for the Creative Commons CC-O license:

```json
{
  "@context": {
    "schema": "https://schema.org",
  },
  "id": "http://www.example-data-repository.org/dataset/123",
  "type": "Dataset",
  "name": "Removal of organic carbon by natural bacterioplankton communities as a function of pCO2 from lab",
  "license": "http://spdx.org/licenses/CC0-1.0"
}
```

**Consequences**

- We gain a comprehensive, maintained, unambiguous vocabulary for licenses, increasing consistency across repositories
- We gain compatibility with the software packaging world like Debian and Python
- Licenses that have well-known URIs (e.g., Creative Commons) may be less recognizable by their SPDX URI
- SPDX license URIs only resolve to HTML pages with machine-readable RDFa embedded, but machine-readable representations in other formats do not seem to be available through content negotiation
Summary Statistics

1,113,210 entities
7,087,380 triples

47,650 Dataset
54,665 DataDownload
599,960 PropertyValue
~ 35k Identifiers
~ 560k Dataset Variables
## 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>omitted ending slash: ‘<a href="http://schema.org%E2%80%99">http://schema.org’</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](https://developers.google.com/search/docs/data-types/dataset)
<table>
<thead>
<tr>
<th>Dataset Properties</th>
<th>Provider Usage</th>
<th>Dataset Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Implemented</td>
<td>Overall</td>
</tr>
<tr>
<td>identifier</td>
<td>30%</td>
<td>100%</td>
</tr>
<tr>
<td>author/creator/contributor</td>
<td>80%</td>
<td>98%</td>
</tr>
<tr>
<td>funder (not awards)</td>
<td>30%</td>
<td>78%</td>
</tr>
<tr>
<td>distribution</td>
<td>60%</td>
<td>100%</td>
</tr>
<tr>
<td>license</td>
<td>70%</td>
<td>98%</td>
</tr>
<tr>
<td>hasPart</td>
<td>10%</td>
<td>2%</td>
</tr>
</tbody>
</table>

“What about Data APIs?”

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

github: earthcubearchitecture-project418/p418Vocabulary
Schema.org Cluster
Common publishing patterns for describing research data on your web pages using schema.org

1. Develop guidelines @ science-on-schema.org

   Telecons:
   - 1st Monday at 5pm ET/2pm PT
   - 4th Thursday at 2:30pm ET/11:30am PT
   more info (google doc)

1. Educate the community through workshops
   ESIP Meetings, AGU, and more...

Tags: Schema.org, Structured Data, RDF

Want to learn more? science-on-schema.org