

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](https://www.w3.org/TR/DWBP)

Biological & Chemical Oceanography Data Management Office

BCO-DMO

Access Data ▾ Submit Data ▾ About Us ▾ Resources ▾ Q

Home / Dataset Search / Southern Ocean 2001 moorings: current data...

Southern Ocean 2001 moorings: current data from ARSV Laurence M. Gould LMG0103, LMG0201A in the Southern Ocean from 2001-2002 (SOGLOBEC project)

Project: U.S. GLOBEC Southern Ocean (SOGLOBEC)

Award # OCE-1851012, Funding by NSF Division of Ocean Sciences

DOI:10.1575/1912/bco-dmo.779540.1 Version Date: Oct. 27 2009 V 1.0 Data type: experimental

VALIDATED

Dataset Files 3 data-filename.CSV (18KB), ISO19115-2.xml (110kb), datapackage.json (5kb)

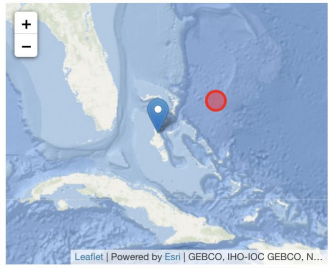
People

Principal Investigator: Robert C Beardsley
Co-Principal Investigator: Dr Richard Limeburner
Student: Dr Carlos Moffat
BCO-DMO Data Manager: Nancy Copley

Abstract

As part of the SO GLOBEC field program, the Woods Hole Oceanographic Institution (WHOI) deployed an array of instrumented subsurface moorings near Marguerite Bay during March 2001- March 2002 and a second array during March 2002-March 2003 (Figure 1). The moored measurements included pressure, temperature, conductivity, velocity, acoustic backscatter, and ice thickness. To monitor surface forcing during the moored array observations, two automatic weather stations (AWSs) were deployed on islands in Marguerite Bay and time series of wind, air temperature, and precipitation were collected from March 2001 through March 2003.

ACCESS DATASET ▾ CITE



Leaflet | Powered by Esri | GEBCO, IHO-IOC GEBCO, N...
N: -66.75003 E: -69.020283 S: -68.25575 W: -70.99985

Iron CTD Profiles CO2

Related Datasets

- Southern Ocean 2001 moorings: current data from ARSV Laurence M. Gould LMG0103,...
- Southern Ocean 2001 moorings: depth and pressure vs. time from ARSV Laurence M. Goul...
- Southern Ocean 2001 moorings: LOW PASS current data from ARSV Laurence M. Gould...

What is Schema.org?

schema:Dataset

schema:title

schema:funding

schema:author

schema:keywords

schema:description

schema:spatialCoverage

- 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](https://www.w3.org/TR/DWBP)

Biological & Chemical Oceanography Data Management Office

Home / Dataset Search / Southern Ocean 2001 moorings: current data...

Southern Ocean 2001 moorings: current data from ARSV Laurence M. Gould LMG0103, LMG0201A in the Southern Ocean from 2001-2002 (SOGLOBEC project)

Project: U.S. GLOBEC Southern Ocean (SOGLOBEC)

Award # OCE-1851012, Funding by NSF Division of Ocean Sciences

DOI:10.1575/1912/bco-dmo.779540.1 Version Date: Oct. 27 2009 V 1.0 Data type: experimental

VALIDATED

Dataset Files 3 data-filename.CSV (18KB), ISO19115-2.xml (110kb), datapackage.json (5kb)

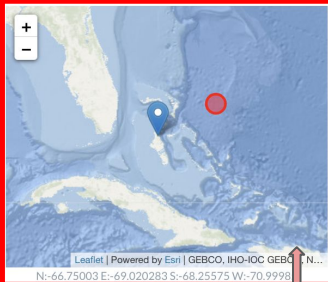
People

Principal Investigator: Robert C Beardsley
Co-Principal Investigator: Dr Richard Limeburner
Student: Dr Carlos Moffat
BCO-DMO Data Manager: Nancy Copley

Abstract

As part of the SO GLOBEC field program, the Woods Hole Oceanographic Institution (WHOI) deployed an array of instrumented subsurface moorings near Marguerite Bay during March 2001- March 2002 and a second array during March 2002-March 2003 (Figure 1). The moored measurements included pressure, temperature, conductivity, velocity, acoustic backscatter, and ice thickness. To monitor surface forcing during the moored array observations, two automatic weather stations (AWSs) were deployed on islands in Marguerite Bay and time series of wind, air

ACCESS DATASET CITE



Leaflet | Powered by Esri | GEBCO, IHO-IOC GEBCO, N...
N:-66.75003 E:-69.020283 S:-68.25575 W:-70.9998

Iron CTD Profiles CO2

Related Datasets

Southern Ocean 2001 moorings: current data from ARSV Laurence M. Gould LMG0103,...

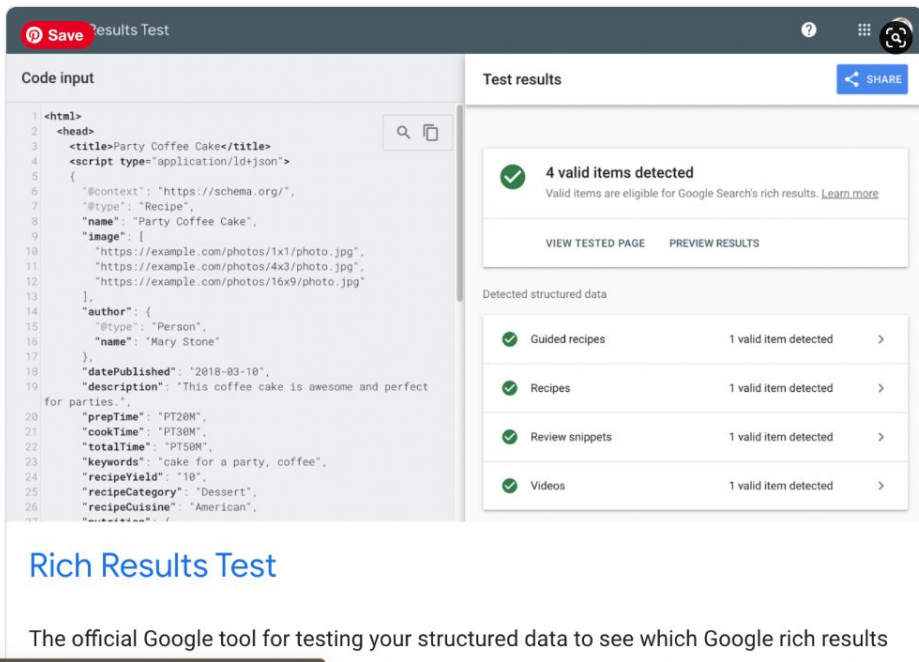
Southern Ocean 2001 moorings: depth and pressure vs. time from ARSV Laurence M. Gould...

Southern Ocean 2001 moorings: LOW PASS current data from ARSV Laurence M. Gould...

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

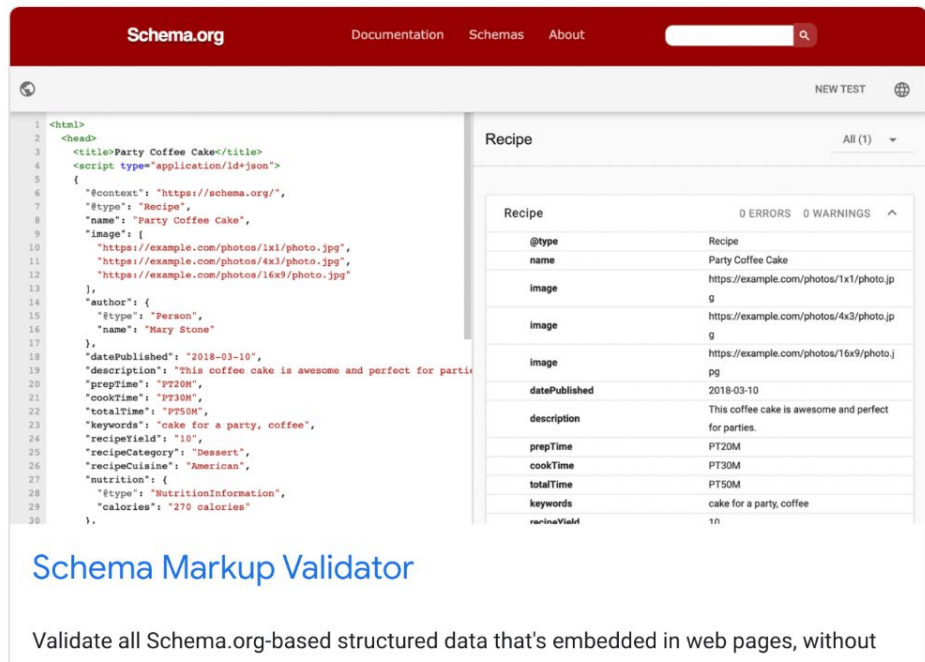
Test Immediately



The screenshot shows the Google Rich Results Test interface. On the left, the 'Code input' section contains a JSON-LD snippet for a 'Party Coffee Cake' recipe. The 'Test results' section on the right shows a green checkmark and the message '4 valid items detected'. Below this, a table lists the detected structured data items:

Detected structured data	Valid items detected	Action
Guided recipes	1 valid item detected	>
Recipes	1 valid item detected	>
Review snippets	1 valid item detected	>
Videos	1 valid item detected	>

Below the table, the 'Rich Results Test' title is followed by the text: 'The official Google tool for testing your structured data to see which Google rich results'.



The screenshot shows the Schema.org Markup Validator interface. The left panel displays the JSON-LD snippet for the 'Party Coffee Cake' recipe. The right panel shows the 'Recipe' schema type with a table of properties and their values:

Property	Value
@type	Recipe
name	Party Coffee Cake
image	https://example.com/photos/1x1/photo.jpg
image	https://example.com/photos/4x3/photo.jpg
image	https://example.com/photos/16x9/photo.jpg
datePublished	2018-03-10
description	This coffee cake is awesome and perfect for parties.
prepTime	PT20M
cookTime	PT30M
totalTime	PT50M
keywords	cake for a party, coffee

Below the table, the 'Schema Markup Validator' title is followed by the text: 'Validate all Schema.org-based structured data that's embedded in web pages, without'.

search.google.com/test/rich-results

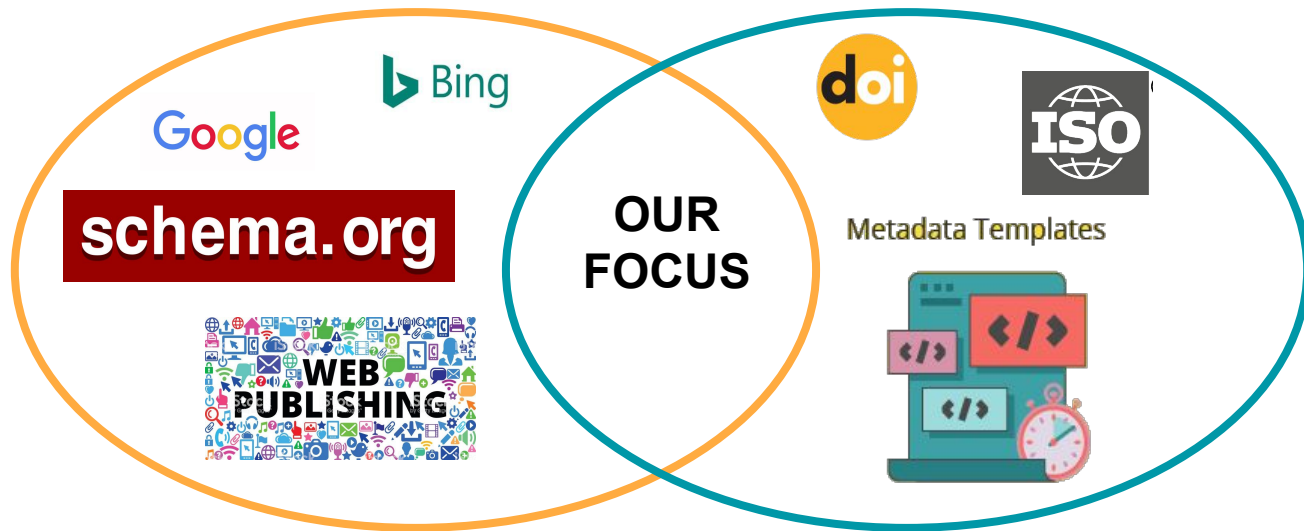
validator.schema.org

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:

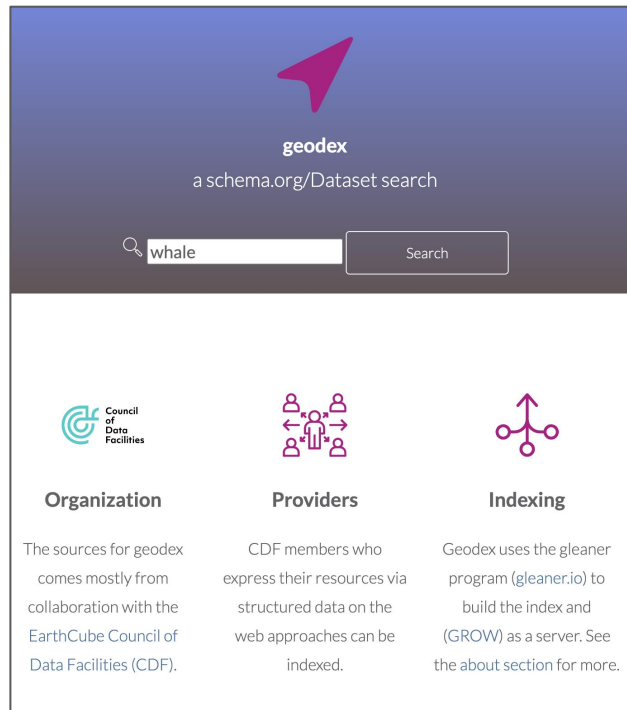
Schema.org, Structured Data, RDF

Want to learn more?

science-on-schema.org

Shared Publishing Patterns

➤ Reliable, consistent ***federation***



The screenshot shows the Geodex website. At the top, there is a purple header with the Geodex logo (a pink arrow) and the text "geodex a schema.org/Dataset search". Below the header is a search bar with the word "whale" and a "Search" button. The main content area is divided into three columns: "Organization" (Council of Data Facilities), "Providers" (CDF members who express their resources via structured data on the web approaches can be indexed), and "Indexing" (Geodex uses the gleaner program (gleaner.io) to build the index and (GROW) as a server. See the about section for more).

➤ Automate Validation

Dataset	Photosymbiosis in planktonic foraminifera across the Palaeocene-Eocene Thermal Maximum
Validation Failed	5 errors. 10 / 23 tests applied.
Violation	Dataset must have an ID
Warning	It is recommended that a Dataset includes a sameAs URL Path: http://schema.org/sameAs
Warning	It is recommended that a Dataset indicates accessibility for free or otherwise Path: http://schema.org/isAccessibleForFree
Violation	Dataset must have a version as Literal or Number Path: http://schema.org/version
Violation	Dataset identifiers must be a URL, Text or PropertyValue Path: http://schema.org/identifier

WHY another Guideline?

<u>variableMeasured</u>	<u>PropertyValue</u> or <u>Text</u>	<u>measurement technique</u> . The variableMeasured property can indicate (repeated as necessary) the variables that are measured in some dataset, either described as text or as pairs of identifier and description using PropertyValue.
Properties from <u>CreativeWork</u>		
<u>about</u>	<u>Thing</u>	The subject matter of the content. Inverse property: <u>subjectOf</u> .
<u>abstract</u>	<u>Text</u>	An abstract is a short description that summarizes a <u>CreativeWork</u> .
<u>accessMode</u>	<u>Text</u>	The human sensory perceptual system or cognitive faculty through which a person may process or perceive information. Expected values include: auditory, tactile, textual, visual, colorDependent, chartOnVisual, chemOnVisual, diagramOnVisual, mathOnVisual, musicOnVisual, textOnVisual.
<u>accessModeSufficient</u>	<u>ItemList</u>	A list of single or combined accessModes that are sufficient to understand all the intellectual content of a resource. Expected values include: auditory, tactile, textual, visual.
<u>accessibilityAPI</u>	<u>Text</u>	Indicates that the resource is compatible with the referenced accessibility API (<u>WebSchemas wiki lists possible values</u>).
	<u>Text</u>	Identifies input methods that are sufficient to fully control the

schema.org

- 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?*

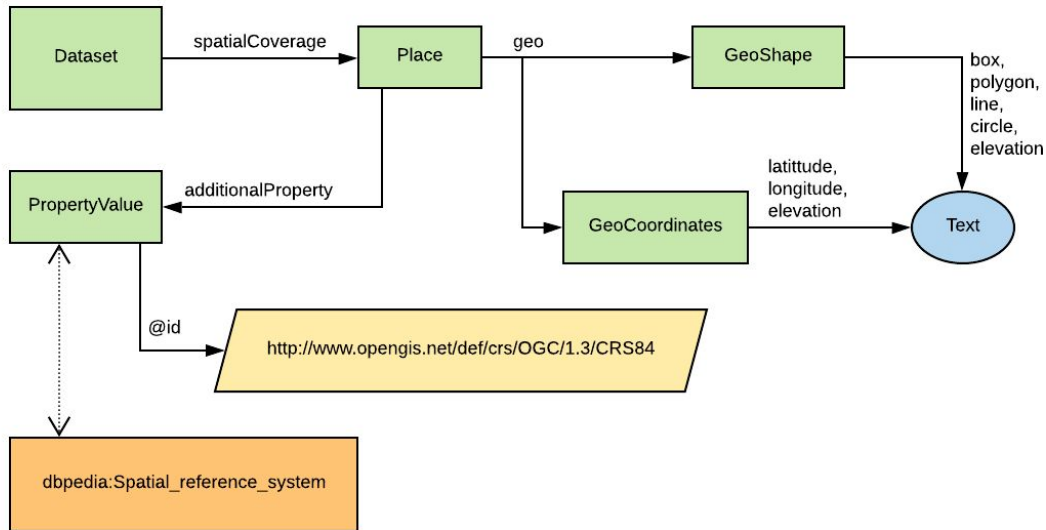
Examples & Drawings

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

```
{
  "@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](#):

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

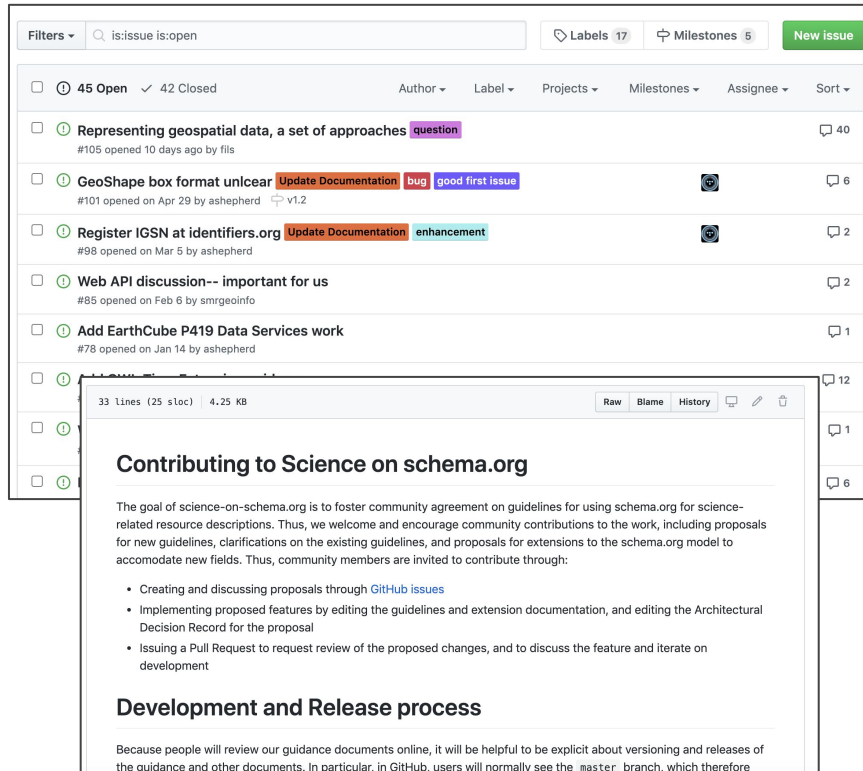


Release Workflow

github.com/.../blob/1.1.0/CONTRIBUTING.md

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'



The screenshot displays the GitHub Issues interface for the 'science-on-schema' repository. At the top, there are filters for '45 Open' and '42 Closed' issues, along with options to filter by labels, milestones, and assignees. A list of issues is shown, including 'Representing geospatial data, a set of approaches', 'GeoShape box format unclear', 'Register IGSN at identifiers.org', 'Web API discussion-- important for us', and 'Add EarthCube P419 Data Services work'. A modal window is open, showing the 'Contributing to Science on schema.org' document. The document outlines the goal of the project, the process for contributing (creating proposals, implementing features, and issuing pull requests), and the development and release process.

Contributing to Science on schema.org

The goal of science-on-schema.org is to foster community agreement on guidelines for using schema.org for science-related resource descriptions. Thus, we welcome and encourage community contributions to the work, including proposals for new guidelines, clarifications on the existing guidelines, and proposals for extensions to the schema.org model to accommodate new fields. Thus, community members are invited to contribute through:

- Creating and discussing proposals through [GitHub issues](#)
- Implementing proposed features by editing the guidelines and extension documentation, and editing the Architectural Decision Record for the proposal
- Issuing a Pull Request to request review of the proposed changes, and to discuss the feature and iterate on development

Development and Release process

Because people will review our guidance documents online, it will be helpful to be explicit about versioning and releases of the guidance and other documents. In particular, in GitHub, users will normally see the `master` branch, which therefore

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

87 lines (66 sloc) 5.02 KB

Raw Blame History

Use SPDX license vocabulary for URIs

Discussion: <https://github.com/ESIPFed/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-0 license:

```
{
  "@context": {
    "@vocab": "https://schema.org/",
  },
  "@id": "http://www.sample-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"
  ...
}
```

readable formats, including
a queryable graph of the

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



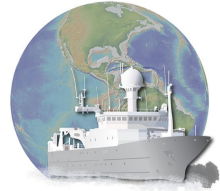
**IALTO
CORDS
GDF RWC**



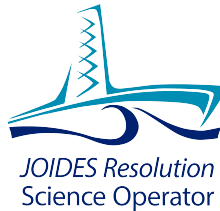
BCO-DMO
Biological & Chemical Oceanography Data Management Office



**Ecoto
tox**

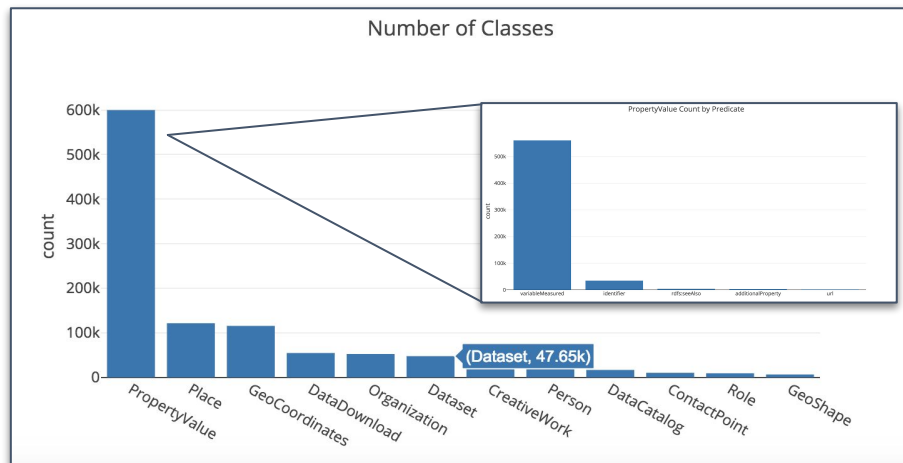
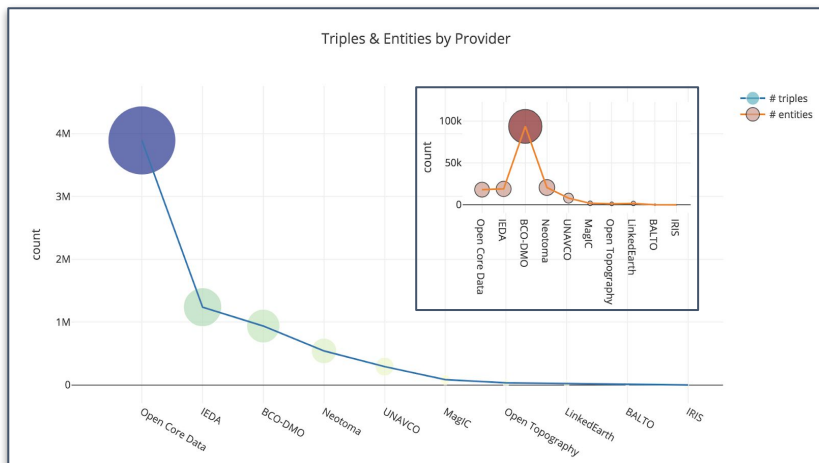


IODP



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

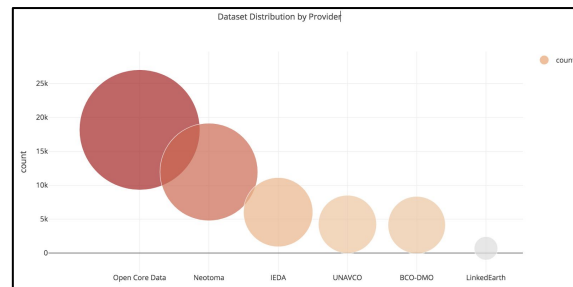
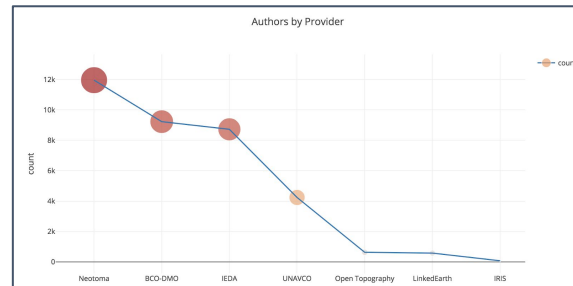
Dataset Properties	Google Requires / Recommends	Provider Usage	Dataset Coverage	
			Implemented	Overall
@context	Required. Set @context to "http://schema.org/"	80%	omitted ending slash: 'http://schema.org'	
@type	Required. Set @type to "Dataset"	100%	47,650 datasets	n/a
name	Required. A descriptive name	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 ex: linking PhysicalSamples to Datasets	10%	122 datasets	2%	<1%

"What about Data APIs?"

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



80 -
100%

50 -
79%

0 -
49%

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