



Dave Vieglais  
Matt Jones

<http://bit.ly/SO-Harvest>

 SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

## Session Overview

---

1. Introduction to schema.org
2. Harvesting workflow
3. Describing datasets
4. Discussion points

Notes: <http://bit.ly/SO-Harvest>

 SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

2

## What is schema.org?

- Collaborative community promoting structured data on the web
- Extensible vocabularies describing entities, actions, relations
- Widely used on millions of sites
- Founded by Google, Microsoft, Yahoo, and Yandex
- Community forum hosted by W3C
  - <https://schema.org/>
  - <https://www.w3.org/community/schemaorg/>
  - <https://github.com/schemaorg/>

## Simplified Content Sharing

### Emergence of schema.org

- Increasing adoption of Schema.org for describing web resources
- Widely adopted embedding of metadata in web pages
- RDF model, typically JSON-LD
- Highly extensible
- Broad coverage vocabulary

*However, application to scientific data still evolving*

[Schema.org/docs/full.html](https://schema.org/docs/full.html)

The screenshot shows the Schema.org full documentation page. The left sidebar lists the class hierarchy under 'Thing':

- Thing
  - Action
    - AchieveAction
      - LoseAction
      - TieAction
      - WinAction
    - AssesAction
      - ChooseAction
        - VoteAction
      - IgnoreAction
      - ReactAction
        - AgreeAction
        - DisagreeAction
        - DislikeAction
  - CreativeWork
    - Article
      - NewsArticle
      - Report
      - ScholarlyArticle
    - SocialMediaPosting
      - BlogPosting
        - LiveBlogPosting
        - DiscussionForumPosting
    - TechArticle
      - APIReference
    - Blog
    - Book
    - Clip
      - MovieClip
      - RadioClip
      - TVClip
      - VideoGameClip

 SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

5

[schema.org/Dataset](https://schema.org/Dataset)

---

Thing > CreativeWork > Dataset

"A body of structured information describing some topic(s) of interest."

Properties:

- distribution
- includedInDataCatalog
- issn
- measurementTechnique*
- variableMeasured*
- + All properties from CreativeWork
- + All properties from Thing

<https://schema.org/Dataset>

The screenshot shows the schema.org/Dataset page. It highlights the inheritance path from Thing to CreativeWork to Dataset. The 'Dataset' class is shown with its subclasses: DataCatalog, Dataset, and DataFeed.

 SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

6

## Extensions

Extending the schema.org vocabularies

Community discussion  
and contributions



- schema.org
- bioschemas.org
- geoschemas.org



**geoschemas.org**

## Schema.org Value

Why add additional information to websites?

- Improve *findability*
  - Google and other indexers

**Google Dataset Search** Beta

🔍

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

[Learn more](#) about including your datasets in Dataset Search.

## Schema.org Value

Why add additional information to websites?

- Improve *findability*
  - Google and other indexers

The screenshot shows the Google Dataset Search interface with the query "weather site:noaa.gov". The results page displays three datasets:

- Radar Weather Observation**: Published Jan 1, 1993. It has a "Explore at data.noaa.gov" button.
- Daily Weather Records**: Published Dec 1, 2013. It has "Explore at catalog-bsp.data.gov" and "Explore at data.wu.ac.at" buttons.
- Mariners Weather Log**: Published 1957. It has "Explore at data.noaa.gov" and "catalog.data.gov" buttons.

Below the results, there is a summary section for the Radar Weather Observation dataset, including links to "View in Google Scholar", "Dataset created Jan 1, 1993", "Dataset published Jan 1, 1993", "Dataset provided by National Oceanic and Atmospheric Administration", "Time period covered Jan 1, 1947 - Jul 31, 2000", and "Area covered Puerto Rico, North Pacific Ocean, Pacific Ocean, United States".

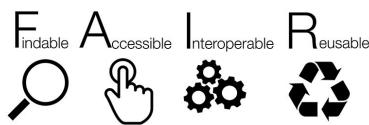
SIP 2019-07-17 bit.ly/SO-Harvest

9

## Schema.org Value

Why add additional information to websites?

- Improve *findability*
  - Google and other indexers
- Facilitate data *access*
  - Programmatic support for resource access
- Enhance *interoperability*
  - Well described resources
- Promote *reuse*
  - Simplified programmatic access



SIP 2019-07-17 bit.ly/SO-Harvest

10

## Exposing schema.org metadata

Easy to add schema.org

Existing Repositories:

1. Take existing web pages, e.g. dataset landing pages
2. Add Schema.org markup (JSON-LD) into <head> section

Or, value added by federations such as DataONE:

1. Schema.org automatically generated at [search.dataone.org](http://search.dataone.org)

 SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

11

## Schema.org Markup

Adding structured knowledge to web pages



The screenshot shows the BCO-DMO website interface. At the top, there's a navigation bar with links for DATA, RESOURCES, ABOUT US, and a search bar. Below the navigation is a sidebar titled "DATABASE" containing a list of categories with their counts: Programs (44), Projects (1,062), Deployments (2,874), Platforms (594), Datasets (9,438), Instruments (487), Parameters (1,419), People (2,702), Affiliations (587), Funding (93), and Awards (1,999). A "Cite This Dataset" button is located in the top right corner of the main content area. The main content area displays a map of the Atlantic Ocean with a red dot indicating a specific location. Below the map, text specifies the spatial extent as N:11.3705 E:-64.519 S:10.05 W:-65.5843 and the temporal extent as 1995-11-03 - 2017-01-12. Project information is listed as CARIACO Ocean Time-Series Program (CARIACO). Principal Investigators listed include Dr Frank Muller-Karger (University of South Florida, USF), Dr Irene Astor (Estacion de Investigaciones Marinas de Margarita, EDIMAR-FLASA), Dr Claudia Benitez-Nelson (University of South Carolina), Dr Mary I. Scranton (Stony Brook University - MSRC, SUNY-SB MSRC), Dr Gordon T. Taylor (Stony Brook University - MSRC, SUNY-SB MSRC), and Dr Robert C. Thunell (University of South Carolina). A "GEOSPATIAL ACCESS" section is also visible.

 SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

12

## Schema.org Markup

Adding structured knowledge to web pages



```

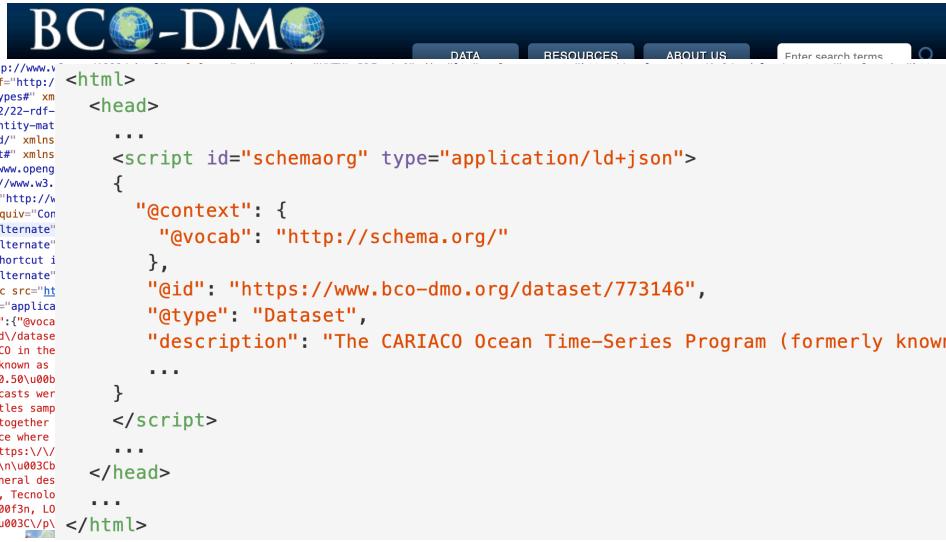
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" version="XHTML-RDFa 1.0" dir="ltr" xmlns:content="http://purl.org/rss/1.0/modules/content/" xmlns:dc="http://purl.org/dc/terms/" xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:og="http://ogp.me/ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:sioc="http://rdfs.org/sioc/ns#" xmlns:sioc="http://rdfs.org/sioc/ns#"; xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:xd="http://www.w3.org/2001/XMLSchema#" xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rss="http://purl.org/rss/1.0/" xmlns:site="https://www.bco-dmo.org/schema/" xmlns:geo="http://ocean-data.org/schema/entity-matching#" xmlns:biblio="http://purl.org/ontology/bibo/" xmlns:crypto="http://id.loc.gov/vocabulary/preservation/cryptographicHashFunctions" xmlns:bcodmo="http://lod.bco-dmo.org/id/" xmlns:datacite="http://purl.org/spar/datacite/" xmlns:arpfo="http://vocab.ox.ac.uk/projectcfunding#" xmlns:tw="http://tw.rpi.edu/schema/" xmlns:dcat="http://www.w3.org/ns/dcat#" xmlns:time="http://www.w3.org/2006/time#" xmlns:geosparql="http://www.opengis.net/ont/geosparql#" xmlns:participation="http://purl.org/vocab/participation/schema#" xmlns:spatial="http://www.w3.org/ns/rdf-spatial#" xmlns:sd="http://www.w3.org/ns/sparql-service-description#" xmlns:dctype="http://purl.org/dc/dcmitype/" xmlns:prov="http://www.w3.org/ns/prov#" xmlns:geolink="http://schema.geolink.org/1.0/base/main#"
  <head profile="http://www.w3.org/1999/xhtml/vocab">
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
    <link rel="alternate" type="application/ld+json" title="JSON-LD Representation" href="https://www.bco-dmo.org/dataset/773146.json">
    <link rel="alternate" type="text/turtle" title="Turtle Representation" href="https://www.bco-dmo.org/dataset/773146.ttl">
    <link rel="shortcut icon" href="https://www.bco-dmo.org/sites/all/themes/bcodmo/images/favicon.ico" type="image/vnd.microsoft.icon">
    <link rel="alternate" type="application/rdf+xml" title="RDF/XML Representation" href="https://www.bco-dmo.org/dataset/773146.rdf">
    <script async src="https://www.bco-dmo.org/sites/default/files/googleanalytics/analytics.js?upphsf"/></script>
  <script type="application/ld+json">
    {"@context": "http://schema.org/", "@id": "https://www.bco-dmo.org/dataset/773146", "identifier": "http://lod.bco-dmo.org/id/dataset/773146", "url": "https://www.bco-dmo.org/dataset/773146", "@type": "Dataset", "name": "CARIACO time series individual CTD profiles from BV/O Hermano Gines H693_CARIACO basin from 1995 (CARIACO project)", "alternateName": "CTD Individual Profiles", "description": "\u00d7003Cpu003ETHE CARIACO Ocean Time-Series Program (formerly known as Carbo Retention In A Colored Ocean) started on November 1995 (CAR-001) and ended on January 2017 (CAR-232). Monthly cruises were conducted to the CARIACO station (10.50\u00b000 N, 64.67\u00b000 W) onboard the RV/Hermano Gines\u00b009s of the Fundaci\u00f3n La Salle de Ciencias Naturales de Venezuela. During each cruise, a minimum of four hydrocasts were performed to collect a suite of core monthly observations. We conducted separate shallow and deep casts to obtain a better vertical resolution of in-situ Niskin-bottles samples for chemical observations, and for productivity, phytoplankton, and pigment observations. One CTD composite profile was created for each cruise by stitching together the sections of the different cruise\u00b700275 CTD profiles at the depth interval where water samples were obtained. CTD\u00b72019s Salinity, Oxygen, and Fluorescence were calibrated with in-situ measurements. The composite CTD profiles dataset is a complement of the hydrographic time series data obtained with the Niskin Bottle Samples (https://www.bco-dmo.org/dataset/7093). The following sections describe the methods used in collecting the core observations at the CARIACO station.\u003Cbr/>\u003E\nMethodology published at CARIACO site (http://imars.usf.edu/publications/methods-cariaco)\u003C/p>\u003E\nAdditional funding support provided by:\u003Cbr/>\u003E\nFondo Nacional de Ciencia, Tecnolog\u00eda e Investigaci\u00f3n FONACIT (2000001782 and 2011000353), Venezuela.\u003Cbr/>\u003E\nInstituto American Institute for Global Change Research, IAI (IAI-CRN3094).\u003C/p>\u003E\n", "isAccessibleForFree": true, "datePublished": "2019-07-16", "keywords": "oceans", "creator": [{"@type": "Person", "@id": "https://www.bco-dmo.org/id/dataset/773146"}]
  </script>

```

SIP 2019-07-17 bit.ly/SO-Harvest 13

## Schema.org Markup

Adding structured knowledge to web pages



```

<html xmlns="http://www.w3.org/1999/xhtml" xmlns:foaf="http://rdfs.org/sioc/types#" xmlns:sioc="http://rdfs.org/sioc/ns#" xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:xd="http://www.w3.org/2001/XMLSchema#" xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rss="http://purl.org/rss/1.0/" xmlns:site="https://www.bco-dmo.org/schema/" xmlns:geo="http://ocean-data.org/schema/entity-matching#" xmlns:biblio="http://purl.org/ontology/bibo/" xmlns:crypto="http://id.loc.gov/vocabulary/preservation/cryptographicHashFunctions" xmlns:bcodmo="http://lod.bco-dmo.org/id/" xmlns:datacite="http://purl.org/spar/datacite/" xmlns:arpfo="http://vocab.ox.ac.uk/projectcfunding#" xmlns:tw="http://tw.rpi.edu/schema/" xmlns:dcat="http://www.w3.org/ns/dcat#" xmlns:time="http://www.w3.org/2006/time#" xmlns:geosparql="http://www.opengis.net/ont/geosparql#" xmlns:participation="http://purl.org/vocab/participation/schema#" xmlns:spatial="http://www.w3.org/ns/rdf-spatial#" xmlns:sd="http://www.w3.org/ns/sparql-service-description#" xmlns:dctype="http://purl.org/dc/dcmitype/" xmlns:prov="http://www.w3.org/ns/prov#" xmlns:geolink="http://schema.geolink.org/1.0/base/main#"
  <head profile="http://www.w3.org/1999/xhtml/vocab">
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
    <link rel="alternate" type="application/ld+json" href="https://www.bco-dmo.org/dataset/773146.json"/>
    <link rel="alternate" type="text/turtle" href="https://www.bco-dmo.org/dataset/773146.ttl"/>
    <link rel="shortcut icon" href="https://www.bco-dmo.org/sites/all/themes/bcodmo/images/favicon.ico" type="image/vnd.microsoft.icon"/>
    <link rel="alternate" type="application/rdf+xml" href="https://www.bco-dmo.org/dataset/773146.rdf"/>
    <script id="schemaorg" type="application/ld+json">
      {"@context": {
        "@vocab": "http://schema.org/"
      },
      "@id": "https://www.bco-dmo.org/dataset/773146",
      "@type": "Dataset",
      "description": "The CARIACO Ocean Time-Series Program (formerly known as Carbo Retention In A Colored Ocean) started on November 1995 (CAR-001) and ended on January 2017 (CAR-232). Monthly cruises were conducted to the CARIACO station (10.50\u00b000 N, 64.67\u00b000 W) onboard the RV/Hermano Gines\u00b009s of the Fundaci\u00f3n La Salle de Ciencias Naturales de Venezuela. During each cruise, a minimum of four hydrocasts were performed to collect a suite of core monthly observations. We conducted separate shallow and deep casts to obtain a better vertical resolution of in-situ Niskin-bottles samples for chemical observations, and for productivity, phytoplankton, and pigment observations. One CTD composite profile was created for each cruise by stitching together the sections of the different cruise\u00b700275 CTD profiles at the depth interval where water samples were obtained. CTD\u00b72019s Salinity, Oxygen, and Fluorescence were calibrated with in-situ measurements. The composite CTD profiles dataset is a complement of the hydrographic time series data obtained with the Niskin Bottle Samples (https://www.bco-dmo.org/dataset/7093). The following sections describe the methods used in collecting the core observations at the CARIACO station.\u003Cbr/>\u003E\nMethodology published at CARIACO site (http://imars.usf.edu/publications/methods-cariaco)\u003C/p>\u003E\nAdditional funding support provided by:\u003Cbr/>\u003E\nFondo Nacional de Ciencia, Tecnolog\u00eda e Investigaci\u00f3n FONACIT (2000001782 and 2011000353), Venezuela.\u003Cbr/>\u003E\nInstituto American Institute for Global Change Research, IAI (IAI-CRN3094).\u003C/p>\u003E\n", "isAccessibleForFree": true, "datePublished": "2019-07-16", "keywords": "oceans", "creator": [{"@type": "Person", "@id": "https://www.bco-dmo.org/id/dataset/773146"}]
    </script>
  </head>
  ...
  <body>
    ...
  </body>

```

SIP 2019-07-17 bit.ly/SO-Harvest 14

## Recommendations

1. Prefer JSON-LD for schema.org markup

## Viewing schema.org

### Some tools

- Web based:
  - [search.google.com/structured-data/testing-tool](https://search.google.com/structured-data/testing-tool)
  - [JSON-LD Playground](https://json-ld.org/playground/)
- Command line / library
  - extract [github.com/scrapinghub/extruct](https://github.com/scrapinghub/extruct)
  - gleaner [github.com/earthcubearchitecture-project418/gleaner](https://github.com/earthcubearchitecture-project418/gleaner)

## Command line, extract

```
5. bash
Last login: Tue Jul 16 09:47:34 on ttys006
(base) $ vieglais$ extract "https://www.archive.arm.gov/metadata/adc/html/nsasondewnpnS01.b1.html" 2>&1 | jq '.["json-ld"][]'
{
  "@context": "https://schema.org",
  "@id": "http://dx.doi.org/10.5439/1021460",
  "@type": "DataSet",
  "about": [
    {
      "@type": "Thing",
      "image": "https://www.arm.gov/img/ARM_Logo.png",
      "name": "Instrument Information: Balloon-borne sounding system (BBSS): Vaisala-processed winds, press., temp, &RH",
      "url": "https://www.arm.gov/capabilities/instruments/sonde"
    },
    {
      "@type": "Thing",
      "image": "https://www.arm.gov/img/ARM_Logo.png",
      "name": "Facility Information: North Slope Alaska",
      "url": "https://www.arm.gov/capabilities/observatories/nsa"
    }
  ],
  "citation": "Atmospheric Radiation Measurement (ARM) user facility. 2015, updated hourly. Balloon-Borne Sounding System (SONDEWNPN). 2015-02-04 to 2019-06-21, North Slope Alaska (NSA) Supplemental 1 (S01). Compiled by D. Holdridge, R. Coulter and J. Kyrouac, ARM Data Center. Data set accessed at http://dx.doi.org/10.5439/1021460.",
  "creator": [
    {
      "@type": "Person",
      "affiliation": {
        "@type": "Organization"
      }
    }
  ]
}
```

 SIP 2019-07-17 bit.ly/SO-Harvest

17

## schema.org harvesting workflow

1. [ robots.txt ]
2. sitemap.xml
3. dataset\_landing\_page.html
4. schema.org structured metadata
5. [ further actions ]





 SIP 2019-07-17 bit.ly/SO-Harvest

18

## schema.org harvesting workflow

Where is the content index?

1. [ robots.txt ]
2. sitemap.xml
3. dataset\_landing\_page.html
4. schema.org structured metadata



User-agent: SomeBot  
Disallow: /

User-agent: \*  
Sitemap: /data/sitemap.xml

- Anyone visiting the domain can easily find the sitemap

## schema.org harvesting workflow

Where is the content?

1. [ robots.txt ]
2. sitemap.xml
3. dataset\_landing\_page.html
4. schema.org structured metadata

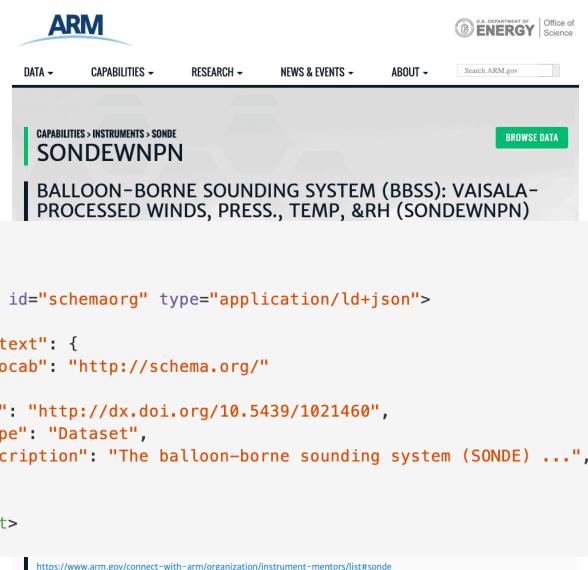


```
<urlset xmlns="http://www.sitemaps.org/schemas/sit
  <url>
    <loc>
      https://www.archive.arm.gov/metadata/adc/html
    </loc>
    <lastmod>2019-06-24</lastmod>
  </url>
  <url>
    <loc>
      https://www.archive.arm.gov/metadata/adc/html
    </loc>
    <lastmod>2019-06-24</lastmod>
  </url>
```

- Anyone visiting the sitemap can easily find the resources

## schema.org harvesting workflow

Resource presented by web server



The screenshot shows the ARM website's header with links for DATA, CAPABILITIES, RESEARCH, NEWS & EVENTS, and ABOUT. Below the header, a breadcrumb navigation shows CAPABILITIES > INSTRUMENTS > SONDE > SONDEWPNP. A green bar highlights the page title "SONDEWPNP". To the right is a search bar and a "BROWSE DATA" button. The main content area displays the schema.org structured metadata for the dataset. The code is as follows:

```

<html>
<head>
...
<script id="schemaorg" type="application/ld+json">
{
  "@context": {
    "@vocab": "http://schema.org/"
  },
  "@id": "http://dx.doi.org/10.5439/1021460",
  "@type": "Dataset",
  "description": "The balloon-borne sounding system (SONDE) ...",
  ...
}
</script>
...

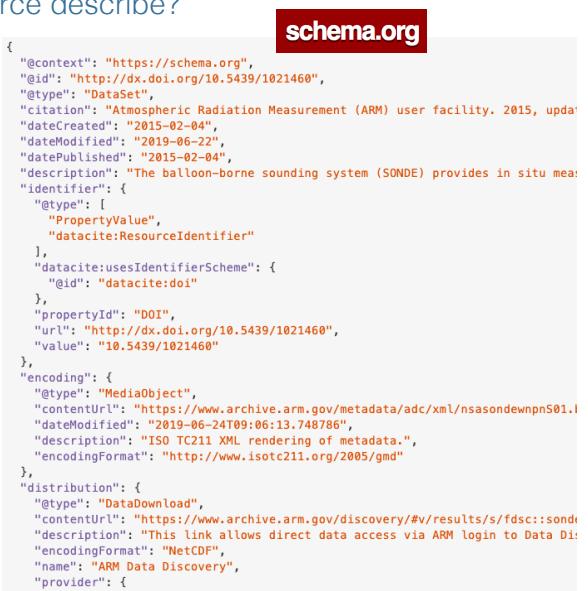
```

<https://www.arm.gov/connect-with-arm/organization/instrument-mentors/list#sonde>

SIP 2019-07-17 bit.ly/SO-Harvest

## schema.org harvesting workflow

What does this resource describe?



The screenshot shows the schema.org structured metadata for the dataset. The code is as follows:

```

{
  "@context": "https://schema.org",
  "@id": "http://dx.doi.org/10.5439/1021460",
  "@type": "DataSet",
  "citation": "Atmospheric Radiation Measurement (ARM) user facility. 2015, update",
  "dateCreated": "2015-02-04",
  "dateModified": "2019-06-22",
  "datePublished": "2015-02-04",
  "description": "The balloon-borne sounding system (SONDE) provides in situ mea...",
  "identifier": {
    "type": [
      "PropertyValue",
      "datacite:ResourceIdentifier"
    ],
    "datacite:usesIdentifierScheme": {
      "@id": "datacite:doi"
    },
    "propertyId": "DOI",
    "url": "http://dx.doi.org/10.5439/1021460",
    "value": "10.5439/1021460"
  },
  "encoding": {
    "@type": "MediaObject",
    "contentUrl": "https://www.archive.arm.gov/metadata/adc/xml/nsasondewnpnS01.i",
    "dateModified": "2019-06-24T09:06:13.748786",
    "description": "ISO TC211 XML rendering of metadata.",
    "encodingFormat": "http://www.isotc211.org/2005/gmd"
  },
  "distribution": {
    "@type": "DataDownload",
    "contentUrl": "https://www.archive.arm.gov/discovery/#v/results/s/fdsc::sonde",
    "description": "This link allows direct data access via ARM login to Data Di...",
    "encodingFormat": "NetCDF",
    "name": "ARM Data Discovery",
    "provider": {
      "@type": "Organization",
      ...
    }
  }
}

```

➤ Structured metadata easily found and parsed

SIP 2019-07-17 bit.ly/SO-Harvest

## schema.org retrieval workflow

Locate resources with minimal starting knowledge

- Progressive introspection
- Discovery of resources with minimal prior knowledge
- Widely applicable pattern for resources on the web
- Tools readily available to generate and consume
- Easy to test and validate

## Recommendations

1. Prefer JSON-LD for schema.org markup
2. Follow widely used practices for web introspection

## schema.org/Dataset

Thing > CreativeWork > Dataset

"A body of structured information describing some topic(s) of interest."

Properties:

- distribution
- includedInDataCatalog
- issn
- *measurementTechnique*
- *variableMeasured*
- + All properties from CreativeWork
- + All properties from Thing

<https://schema.org/Dataset>

[GitHub](#)

[ESIPFed/science-on-schema.org](#)

## Dataset

Simple Dataset

```
{
  "@context": {
    "@vocab": "http://schema.org/"
  },
  "@type": "Dataset",
  "@id": "https://www.sample-data-repository.org/dataset/472032",
  "name": "Removal of organic carbon by natural bacterioplankton communities ...",
  "description": "This dataset includes results of laboratory experiments which ...",
  "url": "https://www.sample-data-repository.org/dataset/472032",
  "sameAs": "https://search.dataone.org/#view/https://www.sample-data-
            repository.org/dataset/472032",
  "version": "2013-11-21",
  "keywords": ["ocean acidification", "Dissolved Organic Carbon", "oceans", ... ],
  "license": "http://creativecommons.org/licenses/by/4.0/",
  "identifier": "urn:sdro:dataset:472032",
}
```

[GitHub](#) [ESIPFed/science-on-schema.org](#)

## Mapping Metadata to Schema.org

Several mappings from existing standards

Schema.org can be generated from existing metadata, e.g.:

- ISO19115
- DCAT
- DublinCore
- DarwinCore

[www.w3.org/2015/spatial/wiki/  
ISO\\_19115\\_-\\_DCAT\\_-\\_Schema.org\\_mapping](http://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping)

Can potentially provide complete replacement of metadata or target a specific role, e.g. for discovery.

## Recommendations

1. Prefer JSON-LD for schema.org markup
2. Follow widely used practices for web introspection
3. Reuse existing identifiers
4. Structured metadata as complete as practicable

## Dataset

Rich metadata already available?

```
{
  "@type": "Dataset", "identifier":{ ... },
  ...
  "encoding":{
    "@type": "MediaObject",
    "contentUrl": "https://example.org/link/to/iso.xml",
    "encodingFormat": "http://www.isotc211.org/2005/gmd",
    "description": "ISO TC211 XML rendering of metadata.",
    "dateModified": "2019-06-12T14:44:15Z"
  }
}
```

## schema.org harvesting workflow

The referenced content

1. [ robots.txt ]
2. sitemap.xml
3. dataset\_landing\_page.html
4. schema.org structured metadata
5. dataset metadata, data, and services

```
<ns0:MD_Metadata xmlns:ns0="http://www.isotc211.org/2005/gmd" xmlns:ns2="http://www.isotc211.org/2005/gmd#"
  xmlns:ns3="http://www.opengis.net/gml" xmlns:xi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.isotc211.org/2005/gmd http://www.isotc211.org/2005/gmd/gmd.xsd">
  <ns0:fileIdentifier>
    <ns2:characterString>nsasasondewnpn801.bi</ns2:CharacterString>
  </ns0:fileIdentifier>
  <ns0:contact>
    <ns0:CI_ResponsibilityParty>
      <ns0:organisationName>
        <ns2:characterString>Atmospheric Radiation Measurement Data Center</ns2:CharacterString>
      </ns0:organisationName>
      <ns0:contactInfo>
        <ns0:CI_Contact>
          <ns0:role>
            <ns0:CI_OnlineResource>
              <ns0:linkage>
                <ns0:URL>https://www.arm.gov</ns0:URL>
              </ns0:linkage>
              <ns0:function>
                <ns0:CI_OnlineFunctionCode>
                  codeListValue="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_OnlineFunctionCode"
                  codeListValueURI="http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI_OnlineFunctionCode"
                </ns0:function>
                <ns0:CI_OnlineResource>
                  <ns0:linkage>
                    <ns0:URL>http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI\_OnlineResource</ns0:URL>
                  </ns0:linkage>
                </ns0:CI_OnlineResource>
              </ns0:role>
              <ns0:role>
                <ns0:CI_OnlineResource>
                  <ns0:linkage>
                    <ns0:URL>http://www.isotc211.org/2005/resources/Codelist/gmxCodelists.xml#CI\_OnlineResource</ns0:URL>
                  </ns0:linkage>
                </ns0:CI_OnlineResource>
              </ns0:role>
            </ns0:CI_ResponsibilityParty>
          </ns0:contact>
          <ns0:dateTimeStamp>
            <ns2:DateTime>2019-06-24T09:06:13.802605</ns2:DateTime>
          </ns0:dateTimeStamp>
          <ns0:citation>
            <ns3:characterString>http://dx.doi.org/10.5439/1021460</ns3:CharacterString>
          </ns0:citation>
          <ns0:identificationInfo>
            <ns0:citation>
              <ns0:CI_Citation>
                <ns0:title>
```



FGDC.GOV  
FEDERAL GEOGRAPHIC DATA COMMITTEE



EML

METS

## Recommendations

1. Prefer JSON-LD for schema.org markup
2. Follow widely used practices for web introspection
3. Reuse existing identifiers
4. Structured metadata as complete as practicable
5. Reference rich external metadata when available

## schema.org/Dataset

Thing > CreativeWork > Dataset

External vocabularies:

- EarthCube Building Blocks – EarthCollab and GeoLink
- Datacite Ontology – DOIs and ORCID
- ViVO Ontology – Datasets



Improve search precision

- Geoscience Standard Names
- GCMD Keywords
- SWEET Ontologies

## Dataset

Well defined identifier

```
{
  "@context": {
    "@vocab": "http://schema.org/",
    "datacite": "http://purl.org/spar/datacite/"
  },
  "@type": "Dataset",
  "url": "https://darchive.mblwholibrary.org/handle/1912/8572",
  "identifier": {
    "@type": ["PropertyValue", "datacite:ResourceIdentifier"],
    "datacite:usesIdentifierScheme": { "@id": "datacite:doi" },
    "propertyID": "DOI",
    "url": "https://doi.org/10.1575/1912/bco-dmo.665253",
    "value": "10.1575/1912/bco-dmo.665253"
  },
  "sameAs": "https://search.dataone.org/#view/https://www.sample-data-
repository.org/dataset/472032",
}
```

 SIP 2019-07-17 bit.ly/SO-Harvest

33

## Recommendations

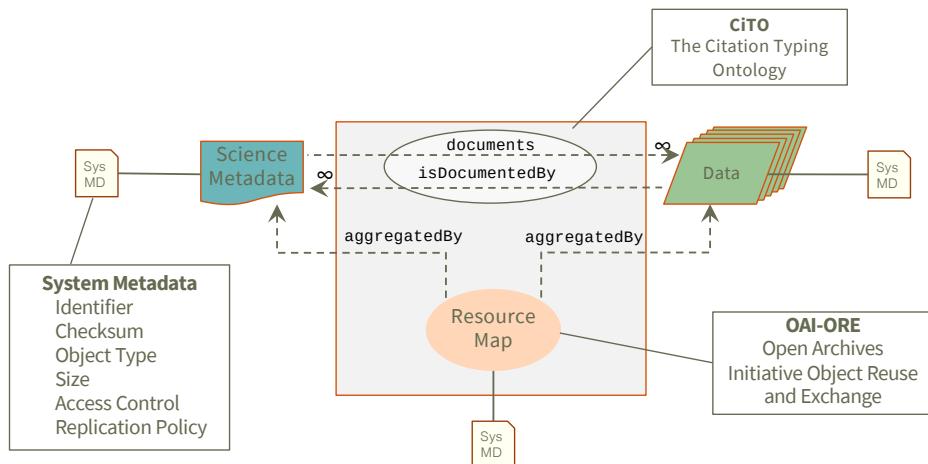
1. Prefer JSON-LD for schema.org markup
2. Follow widely used practices for web introspection
3. Reuse existing identifiers
4. Structured metadata as complete as practicable
5. Reference rich external metadata resource when available
6. Reference well established vocabularies to reduce ambiguity

 SIP 2019-07-17 bit.ly/SO-Harvest

34

## Describing Datasets with Schema.org

### What is a Dataset?



SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

35

## Dataset Model

### OAI-ORE Binding Graph

#### Dataset components

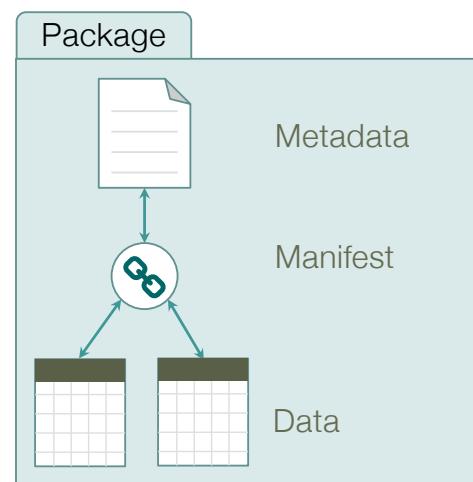
- Metadata

- Data

- Manifest

#### All components

- Immutable
- Uniquely identified
- Resolvable
- Retrievable



SIP 2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

36

# Harvesting

## Dataset knowledge transfer

The screenshot shows the PANGAEA dataset search interface. The top navigation bar includes 'SEARCH', 'SUBMIT', and 'ABOUT'. Below it, a search bar says 'Search for meas' with a magnifying glass icon. A sidebar on the left titled 'Filter by...' contains sections for 'Dataset Author', 'Dataset Publication Year', and 'Topic'. The main content area displays a search result for '97960 datasets found on search in topic Oceans'. Three datasets are listed:

- Tanner, RL; Faye, LE; Stillman, JH (2019):** Respiration, grazing, and defecation rates in *Phyllosiphonia taylori* under crossed salinity/temperature acclimation. Supplement to: Tanner, RL; Faye, LE; Stillman, JH; Temperature and salinity sensitivity of respiration, grazing, and defecation rates in the estuarine eelgrass sea hare, *Phyllosiphonia taylori*. *Marine Biology*. Size: 5 datasets
- Hawkes, NJ; Korabik, M; Beazley, L et al. (2019):** Epibenthic megafauna characteristics in the Emerald Basin during CCGS Hudson cruise HUD-2011-010. Supplement to: Hawkes, NJ; Korabik, M; Beazley, L et al. (2019); Glass sponge grounds on the Scotian Shelf and their associated biodiversity. *Marine Ecology Progress Series*. Size: 2 datasets
- Scheschonk, L; Becker, S; Hehemann, L et al. (2019):** Eco-physiological data on *Laminaria solidungula* and *Saccharina latissima* from Kongsfjorden, Spitsbergen, during the polar night 2016/17. Supplement to: Scheschonk, L; Becker, S; Hehemann, L et al. (2019); Arctic kelp eco-physiology during the polar night in the face of global warming: a crucial role for laminaria. *Marine Ecology Progress Series*. Size: 7 datasets

A diagram showing three overlapping 'Dataset' boxes. Each box contains a central document icon and two smaller grid icons. Arrows point from the first dataset's grid to the second, and from the second's grid to the third, representing the flow of knowledge or data between datasets.

SIP 2019-07-17 bit.ly/SO-Harvest

37

# Harvesting

## Dataset knowledge transfer

The screenshot shows the PANGAEA dataset search interface for 'Oceans'. The top navigation bar includes 'SEARCH', 'SUBMIT', and 'ABOUT'. Below it, a search bar says 'Search for meas' with a magnifying glass icon. A sidebar on the left titled 'Filter by...' contains sections for 'Dataset Author', 'Dataset Publication Year', and 'Topic'. The main content area displays a search result for '97960 datasets found on search in topic Oceans'. Three datasets are listed:

- Tanner, RL; Faye, LE; Stillman, JH (2019):** Respiration, grazing, and defecation rates in *Phyllosiphonia taylori* under crossed salinity/temperature acclimation. Supplement to: Tanner, RL; Faye, LE; Stillman, JH; Temperature and salinity sensitivity of respiration, grazing, and defecation rates in the estuarine eelgrass sea hare, *Phyllosiphonia taylori*. *Marine Biology*. Size: 5 datasets
- Hawkes, NJ; Korabik, M; Beazley, L et al. (2019):** Epibenthic megafauna characteristics in the Emerald Basin during CCGS Hudson cruise HUD-2011-010. Supplement to: Hawkes, NJ; Korabik, M; Beazley, L et al. (2019); Glass sponge grounds on the Scotian Shelf and their associated biodiversity. *Marine Ecology Progress Series*. Size: 2 datasets
- Scheschonk, L; Becker, S; Hehemann, L et al. (2019):** Eco-physiological data on *Laminaria solidungula* and *Saccharina latissima* from Kongsfjorden, Spitsbergen, during the polar night 2016/17. Supplement to: Scheschonk, L; Becker, S; Hehemann, L et al. (2019); Arctic kelp eco-physiology during the polar night in the face of global warming: a crucial role for laminaria. *Marine Ecology Progress Series*. Size: 7 datasets

A diagram showing three overlapping 'Dataset' boxes. Each box contains a central document icon and two smaller grid icons. Arrows point from the first dataset's grid to the second, and from the second's grid to the third, representing the flow of knowledge or data between datasets.

The screenshot shows the DataONE dataset search interface. The top navigation bar includes 'About', 'News', 'Participate', 'Resources', 'Education', 'Data', 'DATAONE SEARCH', 'Search', 'Summary', 'Jump to DOI or ID', and 'Go'. The main content area displays a search result for 'Datasets 1 to 25 of 374,786'. A table lists datasets with columns for ID, Title, and various metrics like size and count. A sidebar on the left titled 'My Search' shows 'Data source: PANGAEA' and a list of filters: Data attribute, Data files, Member Node, Creator, Year, Identifier, Taxon, and Location.

SIP 2019-07-17 bit.ly/SO-Harvest

38

19

## Dataset

Metadata and data

```
{
  "@type": "Dataset", "identifier":{ ... }, "encoding":{ ... },
  "hasPart": [
    {
      "@type": "DataDownload",
      "contentUrl": "https://example.org/link/to/data",
      "encodingFormat": "data format",
      "identifier":{
        "@type": "PropertyValue",
        "propertyId": "UUID",
        "value": "urn:uuid:123e4567-e89b-12d3-a456-426655440000"
      }
    }
  ]
}
```

Documents

Implied Structure

2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

39

## Dataset

Metadata for service

```
{
  "@type": "Dataset", "identifier":{ ... }, "encoding":{ ... },
  "potentialAction": {
    "@type": "SearchAction",
    "target": {
      "@type": "EntryPoint",
      "contentType": ["application/x-netcdf", "text/tab-separated-values"],
      "urlTemplate": "https://www.sample-data-
repository.org/dataset/1234/download?format={format}&startDateTi
me={end}&bounds={bbox}",
      "description": "Download dataset 1234 based on the requested format, start/end dates
and bounding box",
      "httpMethod": ["GET", "POST"]
    },
    "query-input": [
      ...
    ]
}
```

2019-07-17 [bit.ly/SO-Harvest](http://bit.ly/SO-Harvest)

40

## Recommendations

1. Prefer JSON-LD for schema.org markup
2. Follow widely used practices for web introspection
3. Reuse existing identifiers
4. Structured metadata as complete as practicable
5. Reference rich external metadata resource when available
6. Reference well established vocabularies to reduce ambiguity
7. Provide explicit references to dataset components

## Schema.org Summary

- Simple path to adoption
- Lightweight, extensible vocabulary
- Generate from richer standards (ISO19139, EML, etc)
- We can teach Google about data
- Important outside of Google Dataset Search:
  - lingua franca for sharing of structured knowledge on the web
- Evolving: both schema and patterns of use

## Schema.org Challenges

Need for community guidance, best practices

- Identifiers for datasets
- Encouraging identifier reuse
- Dataset packaging
- Modeling of data sets
- Representation of metadata

Example: GitHub ESIPFed/science-on-schema.org



## Discussion Points

- Patterns for dataset representation
- Should all metadata be rendered in schema.org?
- Consistent use of identifiers for Dataset components
- Representing revisions to datasets
- Expressing provenance relationships in schema.org
- Guidelines or best practices when republishing schema.org
- Revise recommendations