METADATA HARVESTING THROUGH SCHEMA.ORG

Dave Vieglais
Matt Jones

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

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

- Thing
  - Action
    - AchieveAction
      - LoseAction
      - TieAction
      - WinAction
    - AssessAction
      - 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

...

- DataCatalog
- Dataset
  - DataFeed
- DigitalDocument
  - NoteDigitalDocument
  - PresentationDigitalDocument
  - SpreadsheetDigitalDocument
  - TextDigitalDocument
- Episode
  - RadioEpisode
  - TVEpisode
- Game
  - VideoGame
- HowTo
  - Recipe
- HowToDirection
- HowToSection
- HowToStep
- HowToTip
- Map
- MediaObject
  - AudioObject
  - DataDownload

schema.org

...

- WebPageElement
  - SiteNavigationElement
  - Table
  - WPAdBlock
  - WPFooter
  - WPHeader
  - WPSideBar
- WebSite

  - Event
    - BusinessEvent
    - ChildrensEvent
    - ComedyEvent
    - CourseInstance
    - DanceEvent
    - DeliveryEvent
    - EducationEvent
    - ExhibitionEvent
    - Festival
    - FoodEvent
    - LiteraryEvent
    - MusicEvent
    - PublicationEvent
      - BroadcastEvent
      - OnDemandEvent
    - SaleEvent
    - ScreeningEvent
    - SocialEvent
    - SportsEvent
    - TheaterEvent
    - VisualArtsEvent

...

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

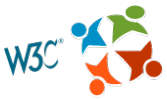
https://schema.org/Dataset

Extensions

Extending the schema.org vocabularies

Community discussion and contributions

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



schema.org

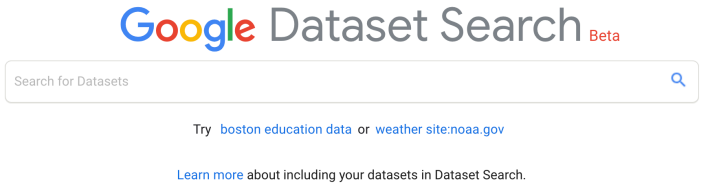
Bioschemas

geoschemas.org

Schema.org Value

Why add additional information to websites?

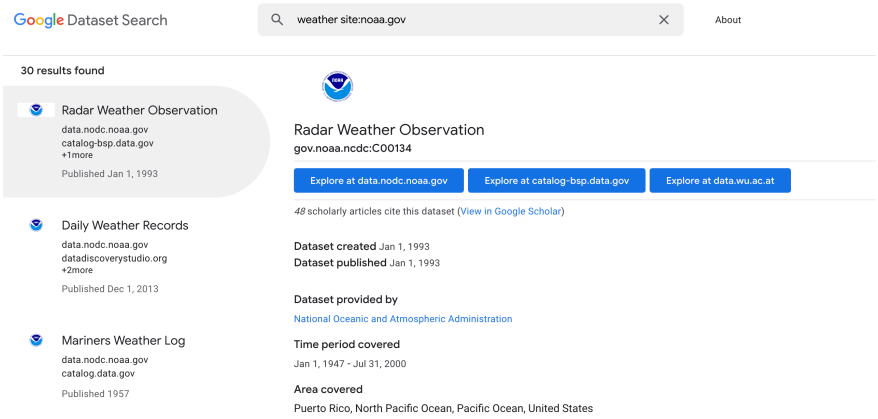
- Improve *findability*
  - Google and other indexers



Schema.org Value

Why add additional information to websites?

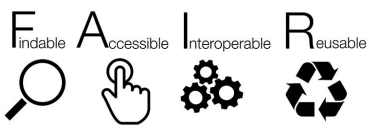
- Improve *findability*
  - Google and other indexers



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



Exposing schema.org metadata

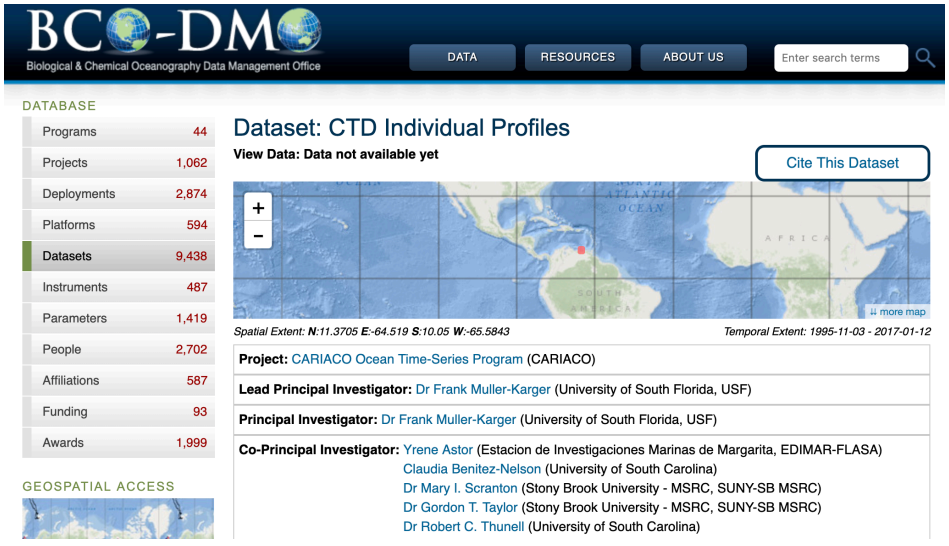
Easy to add schema.org

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

Schema.org Markup

Adding structured knowledge to web pages



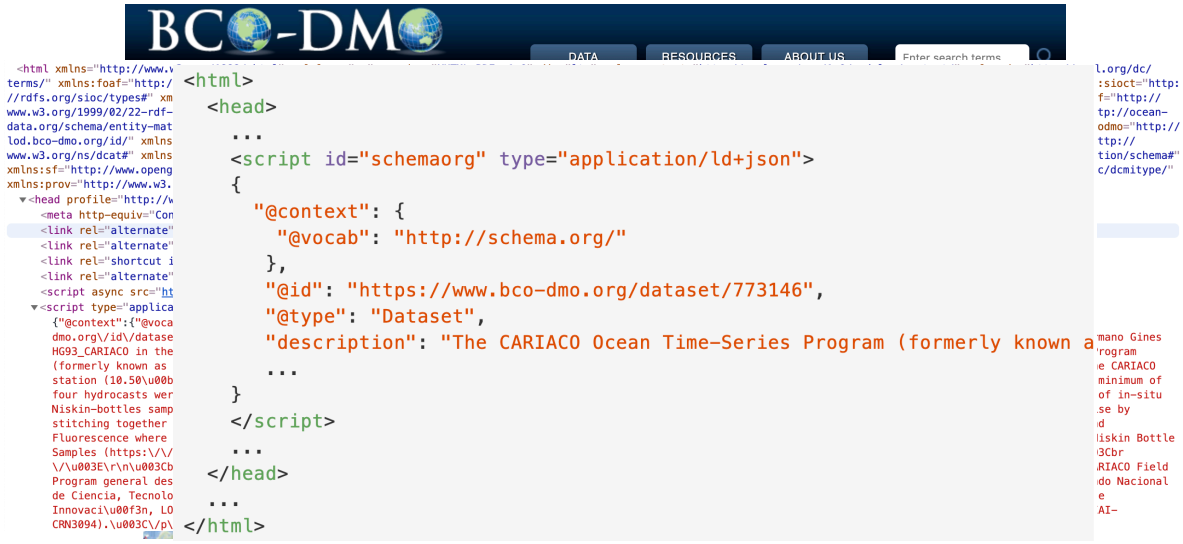
Schema.org Markup

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:sioct="http://rdfs.org/sioc/types#" xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:xsd="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/ns#" xmlns:odo="http://ocean-data.org/schema/" xmlns:emo="http://ocean-data.org/schema/entity-matching#" xmlns:bibo="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/projectfunding#" 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:sf="http://www.opengis.net/ont/sf#" xmlns:void="http://rdfs.org/ns/void#" 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#" class="js">
  ▼<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?puphsf"></script>
    ▼<script type="application/ld+json">
      {"@context":{"@vocab":"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 B\/O Hermano Gines HG93_CARIACO in the CARIACO basin from 1995 (CARIACO project)","alternateName":"CTD Individual Profiles","description":"\u003Cp\u003EThe CARIACO Ocean Time-Series Program (formerly known as CArbon 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\u00b0 N, 64.67\u00b0 W) onboard the R\/V Hermano Gin\u00e9s 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\u0027s CTD profiles at the depth interval where water samples were obtained. CTD\u2019s Salinity, Oxygen, and Fluorescence where 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\/3093). The following sections describe the methods used in collecting the core observations at the CARIACO station.\u003Cbr \/\u003E\r\n\u003Cbr \/\u003E\r\nMethodology published at CARIACO site (http:\/\/imars.usf.edu\/publications\/methods-cariaco)\u003C\/p\u003E\r\n\r\n\u003Cp\u003ECARIACO Field Program general description (http:\/\/www.imars.usf.edu\/cariaco)\u003C\/p\u003E\r\n\r\n\u003Cp\u003EAdditional funding support provided by:\u003Cbr \/\u003E\r\nFondo Nacional de Ciencia, Tecnolog\u00eda e Investigaci\u00f3n, FONACIT (2000001702 and 2011000353), Venezuela.\u003Cbr \/\u003E\r\nLey Org\u00e1nica de Ciencia, Tecnolog\u00eda e Innovaci\u00f3n, LOCTI (Estaci\u00f3n de Investigaciones Marinas, 23914), Venezuela.\u003Cbr \/\u003E\r\nInter-American Institute for Global Change Research, IAI (IAI-CRN3094).\u003C\/p\u003E\r\n","isAccessibleForFree":true,"datePublished":"2019-07-16","keywords":"oceans","creator":[{"@type":"Person","@id":"https:\/\/www.bco-
```


Schema.org Markup

Adding structured knowledge to web pages




```
<html>
  <head>
    ...
    <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 a
      ...
    }
    </script>
    ...
  </head>
  ...
</html>
```

Recommendations

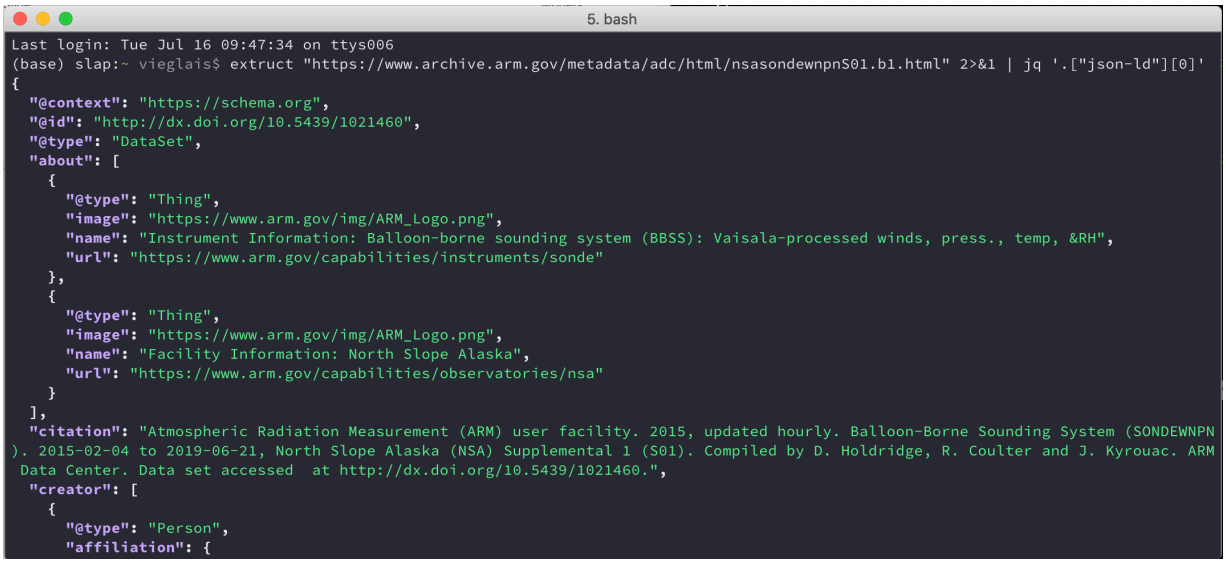
1. Prefer JSON-LD for schema.org markup

Viewing schema.org

Some tools

- Web based:
  - search.google.com/structured-data/testing-tool
  - JSON-LD Playground
- Command line / library
  - extruct github.com/scrapinghub/extruct
  - gleaner github.com/earthcubearchitecture-project418/gleaner

Command line, extruct



schema.org harvesting workflow

1. [robots.txt]
2. sitemap.xml
3. dataset_landing_page.html
4. schema.org structured metadata
5. [further actions]

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

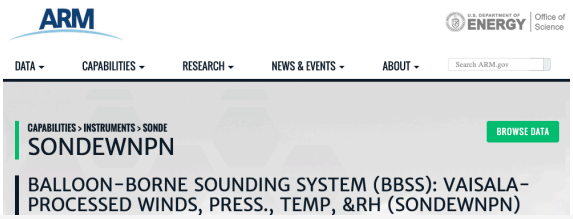
```
▼<urlset xmlns="http://www.sitemaps.org/schemas/sit
  ▼<url>
    ▼<loc>
       https://www.archive.arm.gov/metadata/adc/html
     </loc>
     <lastmod>2019-06-24</lastmod>
   </url>
  ▼<url>
    ▼<loc>
       https://www.archive.arm.gov/metadata/adc/html
     </loc>
     <lastmod>2019-06-24</lastmod>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

➢ Anyone visiting the sitemap can easily find the resources

schema.org harvesting workflow

Resource presented by web server

1. [robots.txt]
2. sitemap.xml
3. dataset_landing_page.html
4. schema.org structured meta




```
<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

schema.org harvesting workflow

What does this resource describe?

1. [robots.txt]
2. sitemap.xml
3. dataset_landing_page.html
4. schema.org structured metadata

➢ Structured metadata easily found and parsed

schema.org


```
{
  "@context": "https://schema.org",
  "@id": "http://dx.doi.org/10.5439/1021460",
  "@type": "DataSet",
  "citation": "Atmospheric Radiation Measurement (ARM) user facility. 2015, upda
  "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.
    "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::sond
    "description": "This link allows direct data access via ARM login to Data Di
    "encodingFormat": "NetCDF",
    "name": "ARM Data Discovery",
    "provider": {
      "@type": "Organization",
```

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

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

ISO/TC 211 FGDC.GOV 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

https://schema.org/Dataset

Dataset

Well defined identifier

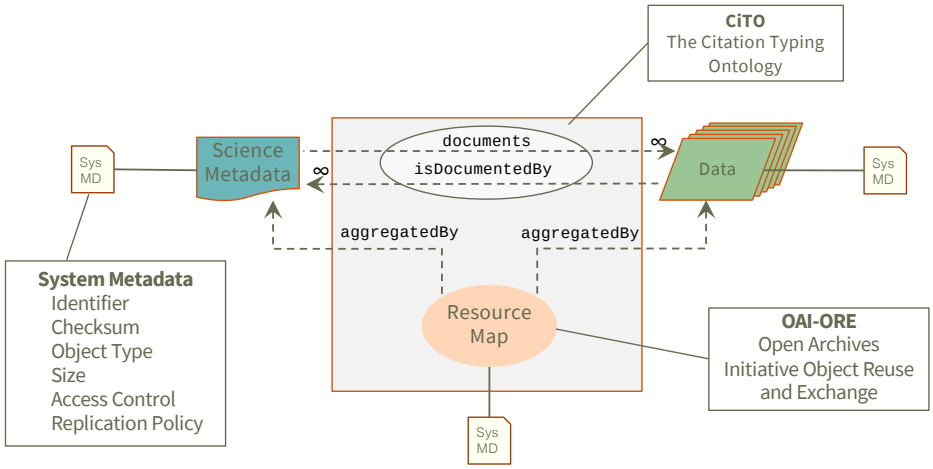

```
{
 "@context": {
   "@vocab": "http://schema.org/",
   "datacite": "http://purl.org/spar/datacite/"
 },
 "@type": "Dataset",
 "url": "https://darchive.mblwhoilibrary.org/handle/1912/8572",
 "identifier": {
   "@type": ["PropertyValue", "datacite:ResourceIdentifier"],
   "datacite:usesIdentifierScheme": {  "@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",
}
```


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

Describing Datasets with Schema.org

What is a Dataset?

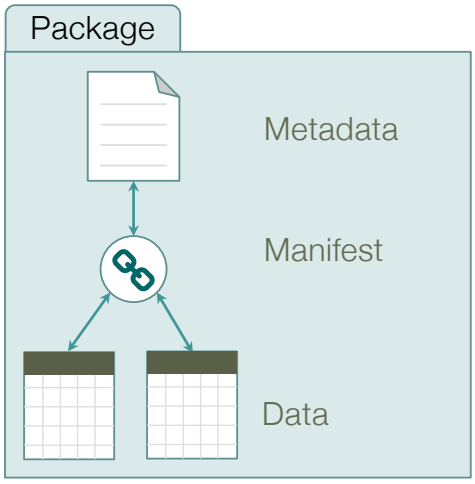


Dataset Model

OAI-ORE Binding Graph

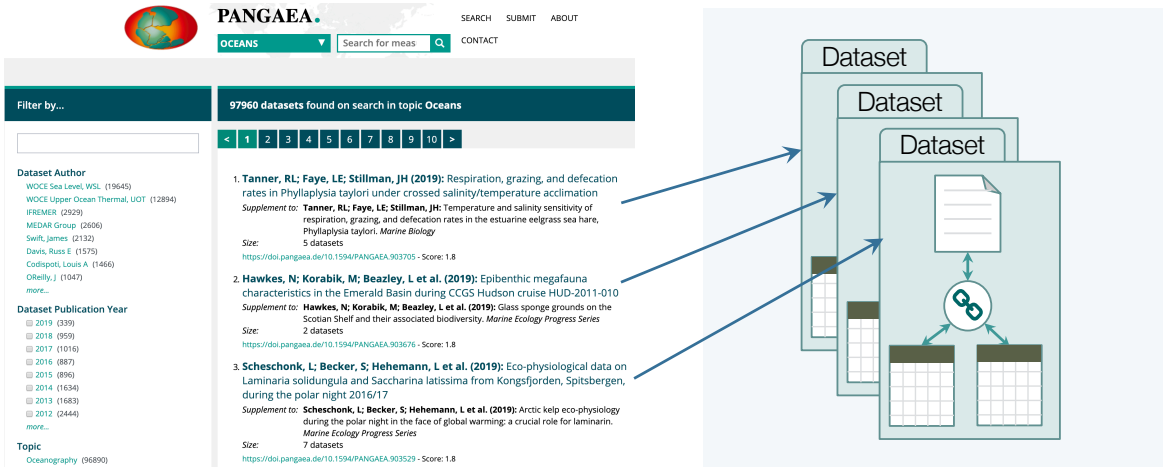
Dataset components
- Metadata
- Data
- Manifest

All components
- Immutable
- Uniquely identified
- Resolvable
- Retrievable



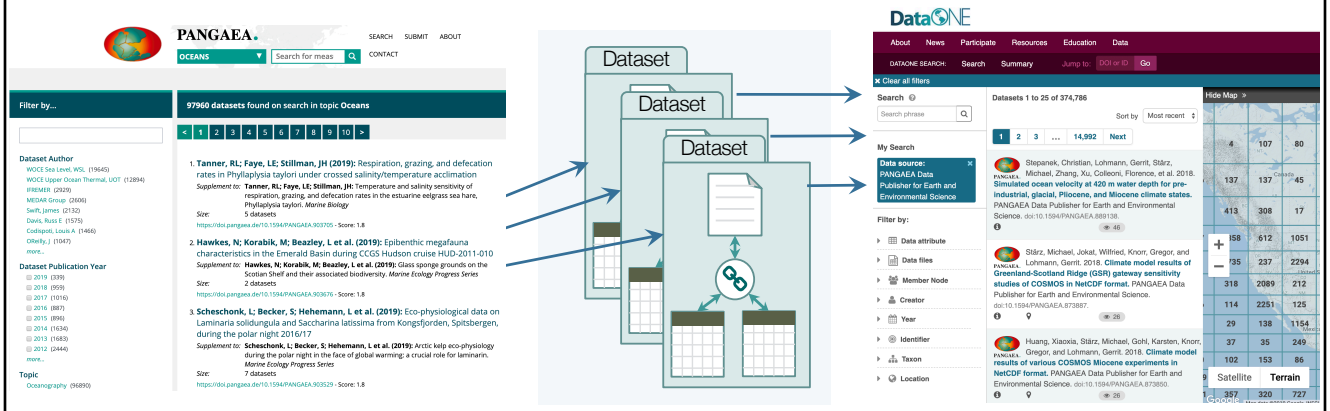
Harvesting

Dataset knowledge transfer



Harvesting

Dataset knowledge transfer



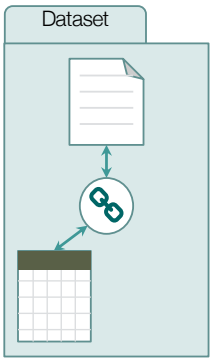
Dataset

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



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}&startDateTime={start}&endDateTi
me={end}&bounds={bbox}",
    "description": "Download dataset 1234 based on the requested format, start/end dates
and bounding box",
    "httpMethod": ["GET", "POST"]
  },
  "query-input": [
  ...

}
```

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