portage

SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

Schema.org for Research Data Managers: A Primer

Chantelle Verhey,
Research Associate, International Technology Office - WDS

Adam Shepherd,
Technical Director, Biological and Chemical Oceanography Data Management Office (BCO-DMO)

WORLD
DATA SYSTEM

Outline

1. **Introduction**
2. **Schema.org**
   - What is it? How does it work?
   - History
3. **Benefits**
   - Driving Factors (i.e. Google Dataset Search(GDSS))
   - Workflows
   - Interoperability
4. **Considerations**
   - Lightweight; Vocabulary Alignment
5. **Crosswalks**
   - Research Data Alliance (RDA) Research Metadata Schemas, Working Group (WG) **Crosswalk Visualizations**
6. **Domain rich extensions**
   - Science-on-schema.org

WORLD DATA SYSTEM

Introduction

Data managers are at the vanguard of the open data movement, driven by the desire to ensure that our data assets are well managed and widely available to the research community and the public more broadly.

- encapsulated in the **FAIR** principles for data, and the **TRUST** principles (Lin, et al., 2020) for data repositories.
- **FAIR** principles promote activities that ensure data is **F**inable, **A**ccessible, **I**nteroperable and **R**eusable,
- **TRUST** principles of **T**ransparency, **R**esponsibility, **U**ser focus, **S**ustainability and **T**echnology.
- Semantic markup is a core technology we have at our disposal to ensure that data managers adhere to these principles.

What is Schema.org?

Schema.org (pronounced "schema dot org" or SDO)

is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

A shared vocabulary makes it easier for webmasters and developers to decide on a schema and get the maximum benefit for their efforts.

The schemas are a set of 'types', each associated with a set of properties. The types are arranged in a hierarchy.

- The vocabulary currently consists of 778 Types, 1383 Properties 15 Datatypes, 73 Enumerations and 367 Enumeration members.

WORLD
DATA SYSTEM

History

The idea of the semantic web long predates the arrival of SDO.

- Tim Berners-Lee (1999), the inventor of the World Wide Web, expressed a vision of the web as a connected set of data.

The goal of the semantic web is to make content on the web machine readable and actionable by making explicit connections between published content.

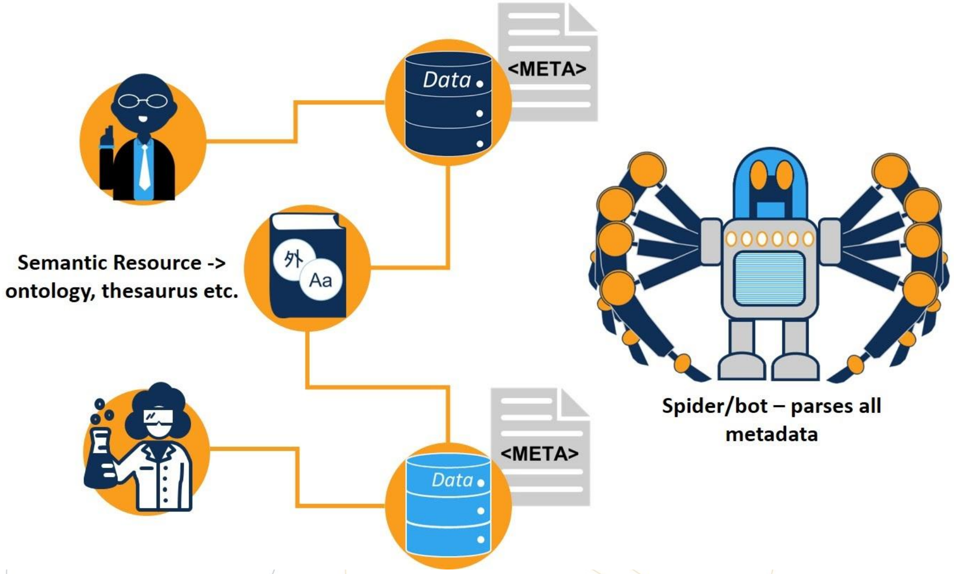
- As data managers and scientists with increasing volumes and variety of data to manage, discover and analyze, it is imperative that we adopt semantic markup and related technologies so that we can automate our workflows.
- SDO is a specific example of the broader class of semantic technologies. SDO was created by four major search engines in 2011: Google, Microsoft, Yahoo, Bing and Yandex (Guha, 2011).

Benefits

- Google Dataset Search (GDSS) is the primary driver for repositories implementing SDO in their metadata landing pages.
- GDSS was designed to do for datasets what Google Scholar (GS) did for publications,
    - providing a single interface to search and view simple metadata records describing datasets from thousands of data repositories around the world.
- The natural appeal of both GDSS and GS stems from the user friendly interface that Google creates, making it an easy go to for finding content (Gusenbaur, 2019).
- Once a data repository has implemented SDO in their dataset landing pages, they can become discoverable in GDSS.
    - At this time Google does not hold copies of the data itself. Users who discover data in GDSS are referred back to the repository to get access to the data.

Workflow

- Workflow is this:
  - Repositories have a database of metadata that describe their holdings.
  - That set of metadata is used to either publish a harvestable metadata service and/or is used to generate a series of landing pages, one per dataset.
  - That landing page is a web page generated from the database of metadata and includes a pointer to the data resource that can be found at the repository.
  - **The landing pages are marked up with SDO terms and indexed and included in GDSS.**
  - Importantly for those who manage sensitive or restricted data which are subject to privacy concerns, once a dataset is discovered in GDSS the user is directed back to the repository to obtain the data.
  - The access restrictions and protocols for data sharing, as agreed upon during the submission process are maintained at the repository.


Data
<META>
Semantic Resource -> ontology, thesaurus etc.
外
Aa
Spider/bot – parses all metadata
Data
<META>

The basic premise of the semantic web is that one person publishes content, and the terms used to describe that content are defined in an online semantic resource (a dictionary, an ontology or a list of controlled terms).

The semantic resource must be online and open to parsing by machines. For data managers, this opens the possibility of machines being able to identify, compare, conflate and process data.

- In our case, Google spiders can crawl over a set of dataset landing pages, collect all the titles, abstracts and location information for each dataset, index them and open that index up through GDSS for search.

Interoperability

There are only 2 terms that are required as part of a *dataset* description to be included in the GDSS index: name and description.

- In addition, there are *additional terms that Google recommends* you include as part of your metadata markup, including but not limited to the dataset creator, download information, unique identifier, license and variables measured.

SDO supports three different mechanisms or formats to add vocabulary terms to a webpage: microdata, RDFa and JSON-LD.

- Both microdata and RDFa are a set of tags and HTML5 extensions that are embedded inline in a webpage next to the text it defines.

WORLD DATA SYSTEM

Considerations

Aligning Semantic Markup Across Domains

For data managers, SDO serves at least 3 functions.

1) It immediately opens our data holdings to a wider audience via GDSS.

2) It is a high level, simple introduction to semantic markup and serves as a gateway to more complex ontologies and workflows.

3) It is a good starting point for aligning ontologies (and by extension our datasets) across domains.

WORLD DATA SYSTEM

Aligning Semantic Markup Across Domains

- Data managers often use multiple ontologies to annotate or mark up documents like metadata.
- Many hope that they can use a handful of common terms from SDO (like "title," and "description"), and reference other ontologies to capture more domain specific knowledge about their datasets.
- To the extent possible, even when developing domain specific ontologies, the best practice is to re-use terms from well established, and well served vocabularies.
    - For example, the World Meteorological Organization has a *controlled vocabulary* that the ocean and polar communities typically use called the *International Meteorological Vocabulary* and is translated into four languages (WMO, International Meteorological Vocabulary, 1992).

WORLD
DATA SYSTEM

Although…

- The creation of controlled domain specific vocabularies is fairly labour intensive and requires quite a bit of collaboration from within scientific communities to reach a consensus on which controlled vocabularies should be included in the extension and how they should be implemented (Jonquet, J., et al, 2018).
- The important message here is to avoid reinventing the wheel; if domains can look for guidance from other communities that already have completed this step
  - It is important to recognize large initiatives like the robust set of Paleoenvironmental Standard Terms (PaST) from the Paleoclimatology community and the Medical Subject Headings (MeSH) from the National Library of Medicine are important pillars in scientific semantics.

WORLD
DATA SYSTEM

RDA Research Metadata Schemas WG

Output 1: Data Model

- A generic ‘conceptual data model’ with essential types and properties for research data discovery over the web.

Output 2: Producing a Guidance Document

- A guideline, illustrated with common patterns, of common patterns for publishing metadata landing pages with structured data markups; and a guideline of how to customize the research schemas for target domains with examples.

Output 3: Toolings

- Toolings for making the implementation easier if resources are available.

Welcome to Schema Crosswalks

Visualizations of Schema Crosswalks

Roadmap for:

A. WG outputs

B. **Visualizations**

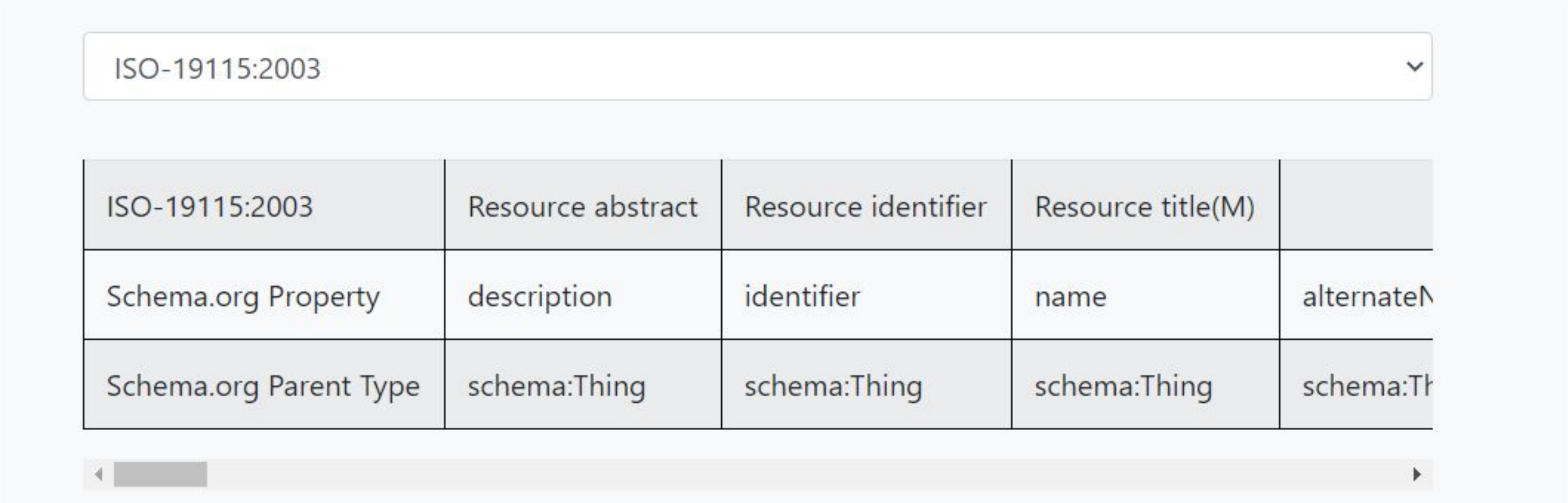
- ~15 Metadata Schemas mapped
- Crosswalks currently included: DCAT-AP,DCATv2, Datacite, ISO19115, EOSC/EDMI, Dataverse, DATS, RIF-CS, DC, Bioschema, B2Find, DDI, ECRIN, CODEMETA, SPASE

https://rd-alliance.github.io/Research-Metadata-Schemas-WG/

WORLD DATA SYSTEM

B. Visualizations – Filter Table

- Schema Filter table provides a 1-to-1 mapping of metadata terms to schema.org
- Utilizes the top level of Schema.org
- Identifies any missing terms (from the schema or from SDO suggested terms) with no currently mapping

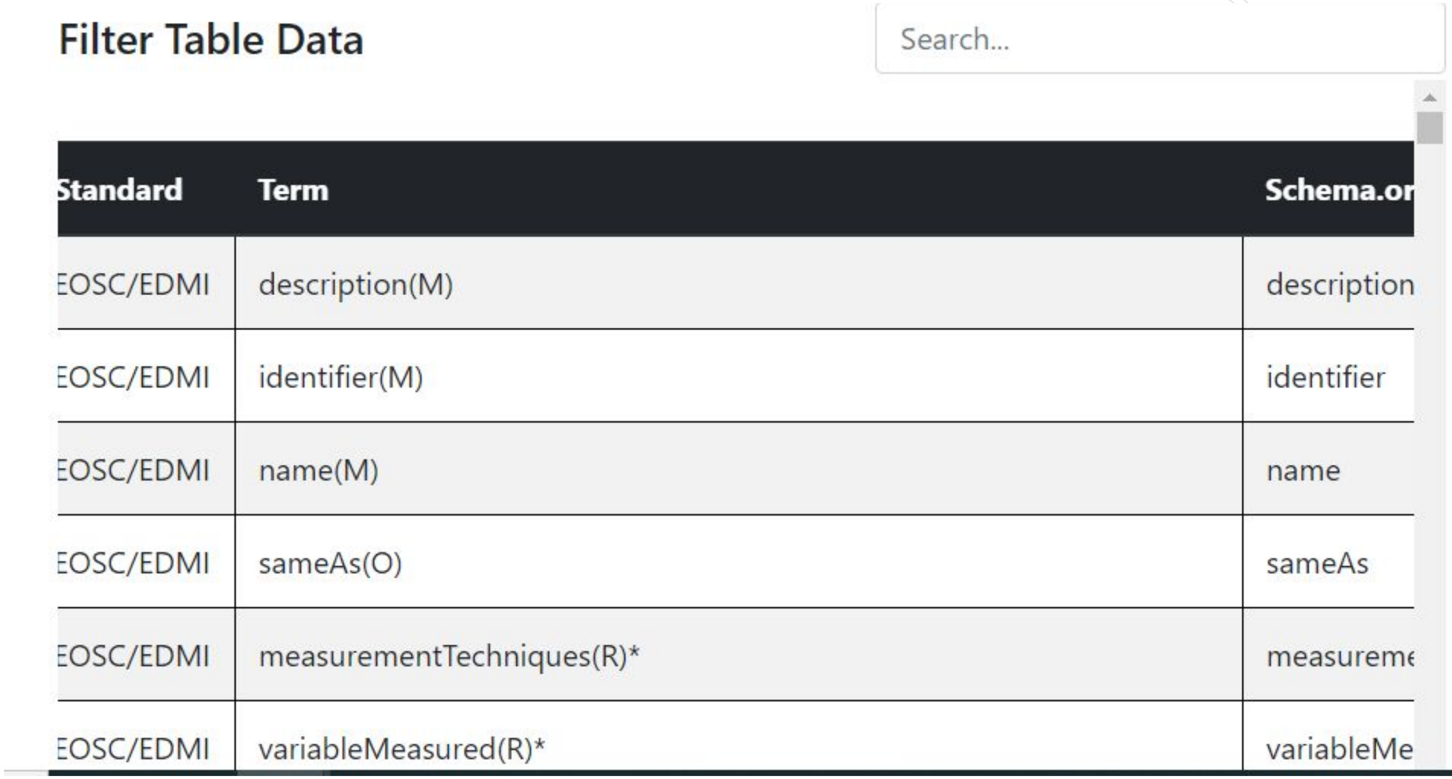


ISO-19115:2003

ISO-19115:2003	Resource abstract	Resource identifier	Resource title(M)	
Schema.org Property	description	identifier	name	alternateN
Schema.org Parent Type	schema:Thing	schema:Thing	schema:Thing	schema:Th

B. Visualizations – Search Table

- Table is a free text search over both metadata and schema.org properties
- Table will pull all related metadata terms for associated property searched
- For example: a search for “publish” will not return records, but the search for ‘publish’ will return ‘datePublished”, “publisher”, “Dataset Publisher”



Filter Table Data

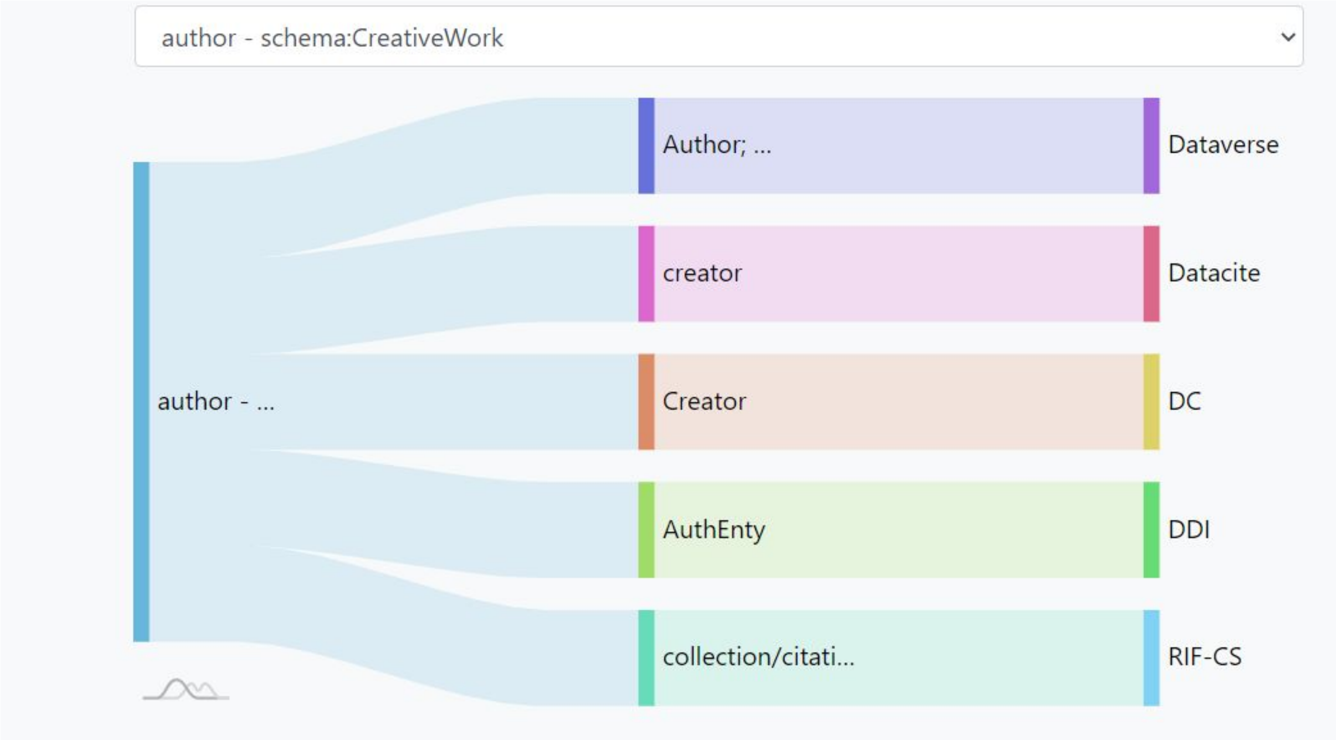
Standard	Term	Schema.or
EOSC/EDMI	description(M)	description
EOSC/EDMI	identifier(M)	identifier
EOSC/EDMI	name(M)	name
EOSC/EDMI	sameAs(O)	sameAs
EOSC/EDMI	measurementTechniques(R)*	measureme
EOSC/EDMI	variableMeasured(R)*	variableMe

B. Visualizations – Sankey Diagram

- This filter allows you to choose a schema.org property and see which (if any) crosswalk term is connected to which standard.

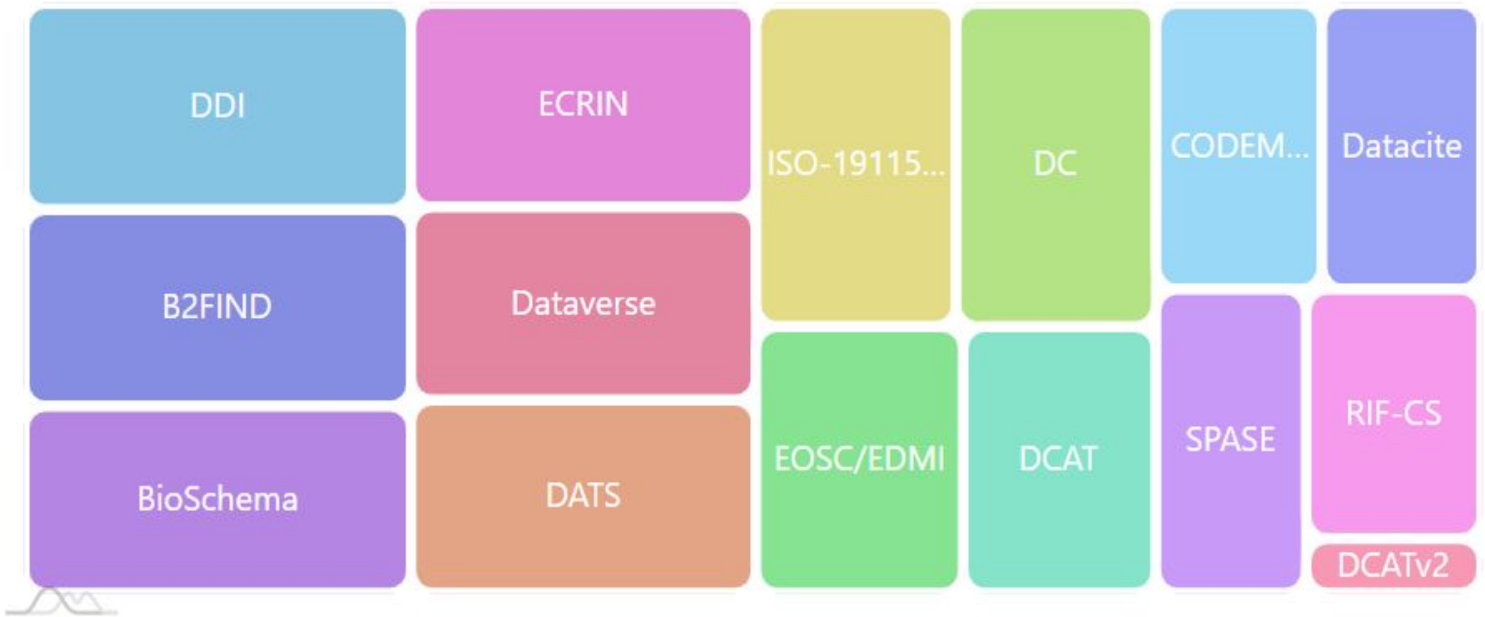
From left to right labels go:

Schema.org properties Crosswalk Term Metadata standard



B. Visualizations – Gap Analysis

- Gap Analysis gives a visual representation of terms within the metadata standards that did not have a schema.org property.
- By hovering over the metadata standard name, it will display a numerical value



Feedback for Visualizations

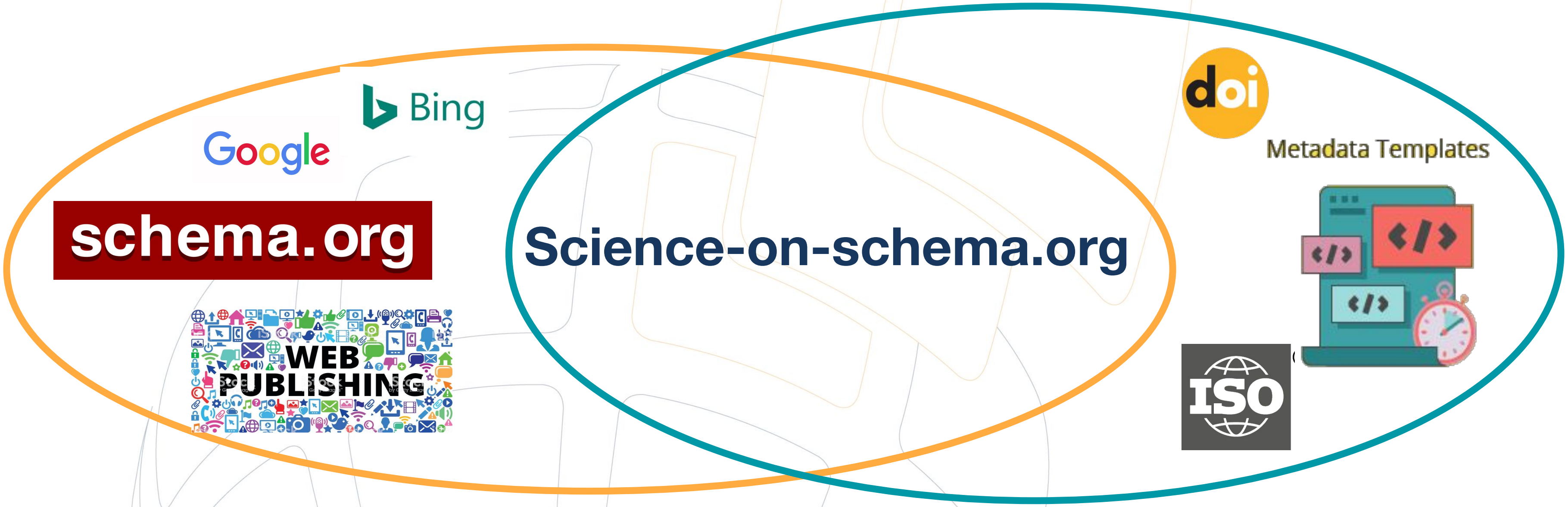
RDA Plenary 16 – Feedback

- Gap Analysis: Include commonly missed terms/ analysis from SDO and/or schemas

- Sankey Diagram: Pop-out window to replace the ‘hover’ feature

- The current PLAN for the home for the RDA Research Metadata Schemas Working Groups crosswalks:
  - Zenodo
  - FAIRsharing

Domain Extensions

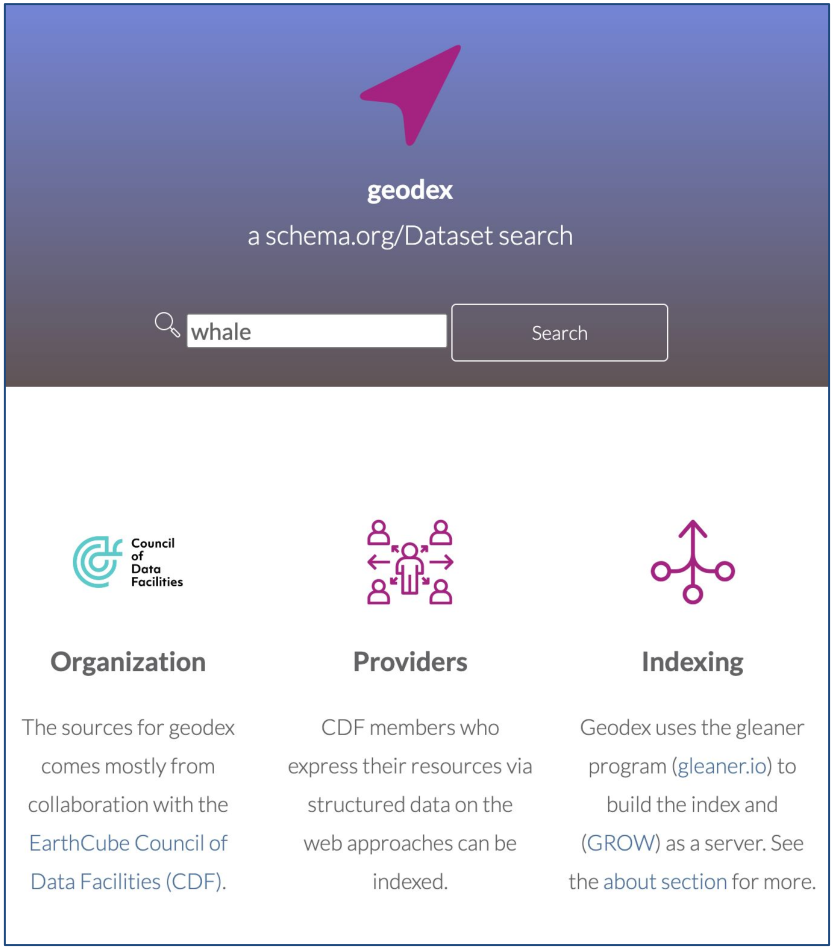


Science-on-schema.org

Shared publishing <u>patterns</u> for describing research data using *schema.org*

Shared Publishing Patterns

- ➢ Reliable, consistent ***federation***
- ➢ Automate Validation



geodex

a schema.org/Dataset search

Organization

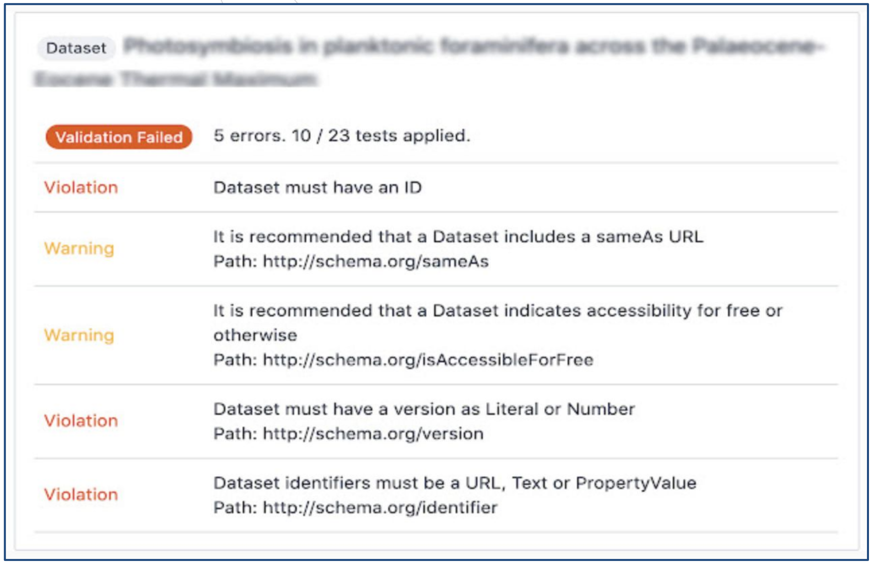
The sources for geodex comes mostly from collaboration with the EarthCube Council of Data Facilities (CDF).

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.

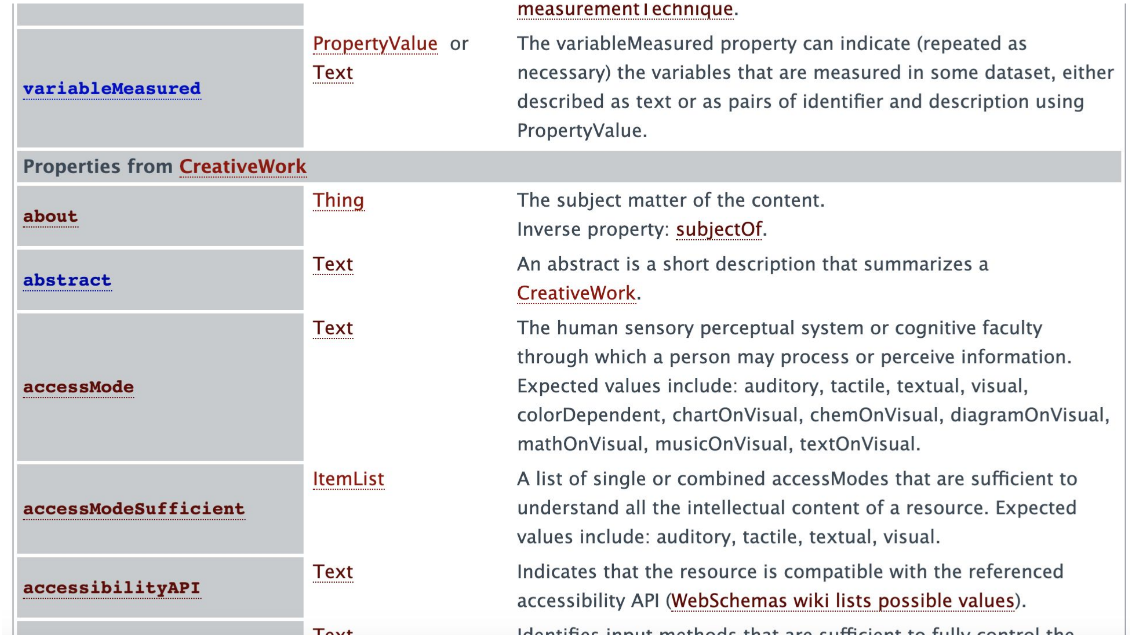


Dataset Photosymbiosis in planktonic foraminifera across the Palaeocene-Eocene Thermal Maximum

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

WORLD DATA SYSTEM

Why another Guideline?



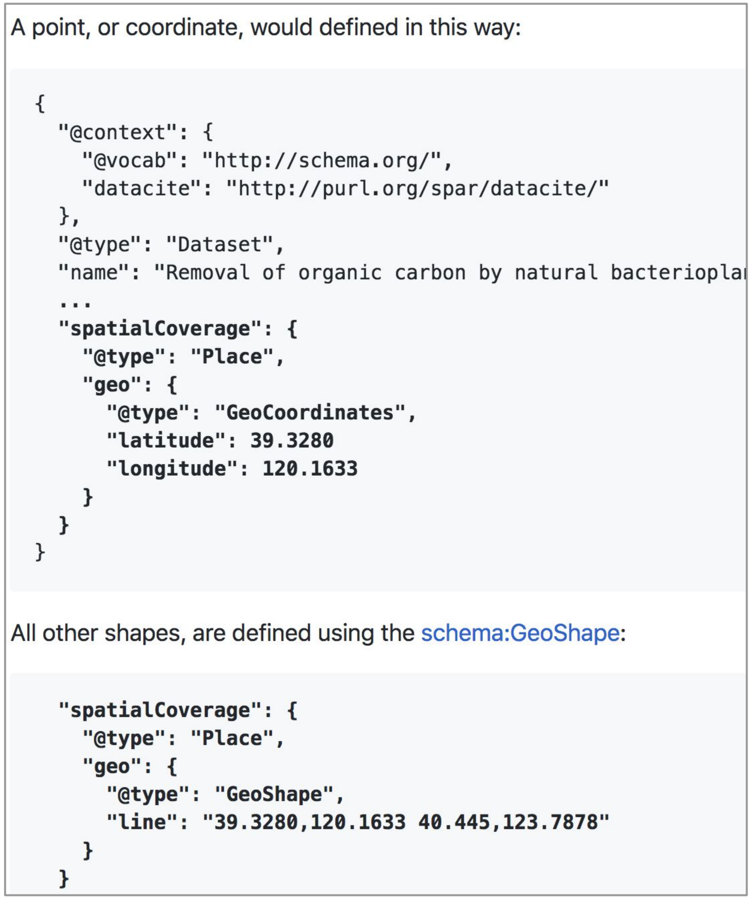
		measurementTechnique.
variableMeasured	PropertyValue or Text	The variableMeasured property can indicate (repeated as necessary) the variables that are measured in some dataset, either described as text or as pairs of identifier and description using PropertyValue.
Properties from CreativeWork		
about	Thing	The subject matter of the content. Inverse property: subjectOf.
abstract	Text	An abstract is a short description that summarizes a CreativeWork.
accessMode	Text	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.
accessModeSufficient	ItemList	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.
accessibilityAPI	Text	Indicates that the resource is compatible with the referenced accessibility API (WebSchemas wiki lists possible values).
	Text	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?***

Code Samples & Drawings

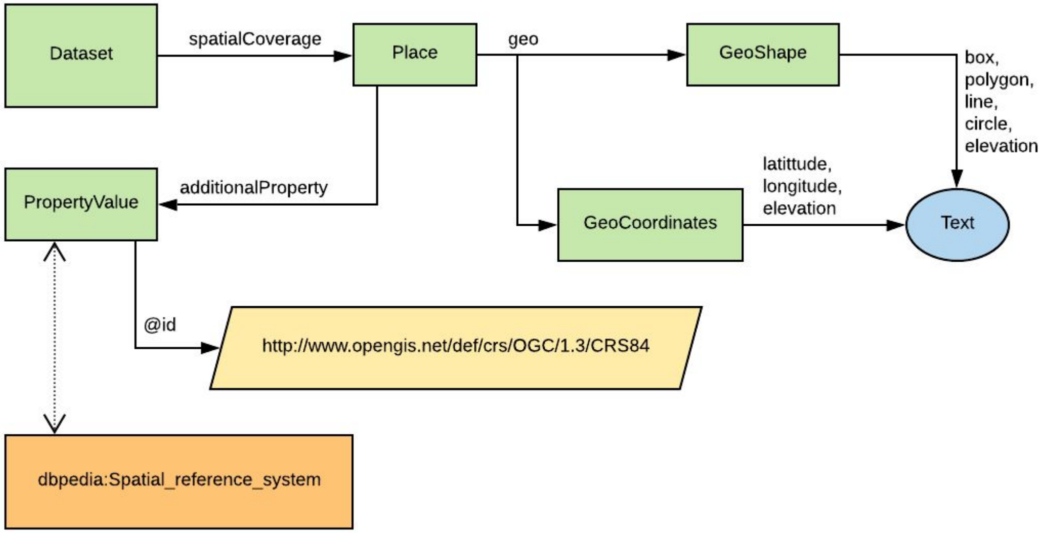


A point, or coordinate, would defined in this way:

```
{
  "@context": {
    "@vocab": "http://schema.org/",
    "datacite": "http://purl.org/spar/datacite/"
  },
  "@type": "Dataset",
  "name": "Removal of organic carbon by natural bacteriopla
  ...
  "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"
    }
  }
```



WORLD DATA SYSTEM

Latest Release: v1.1

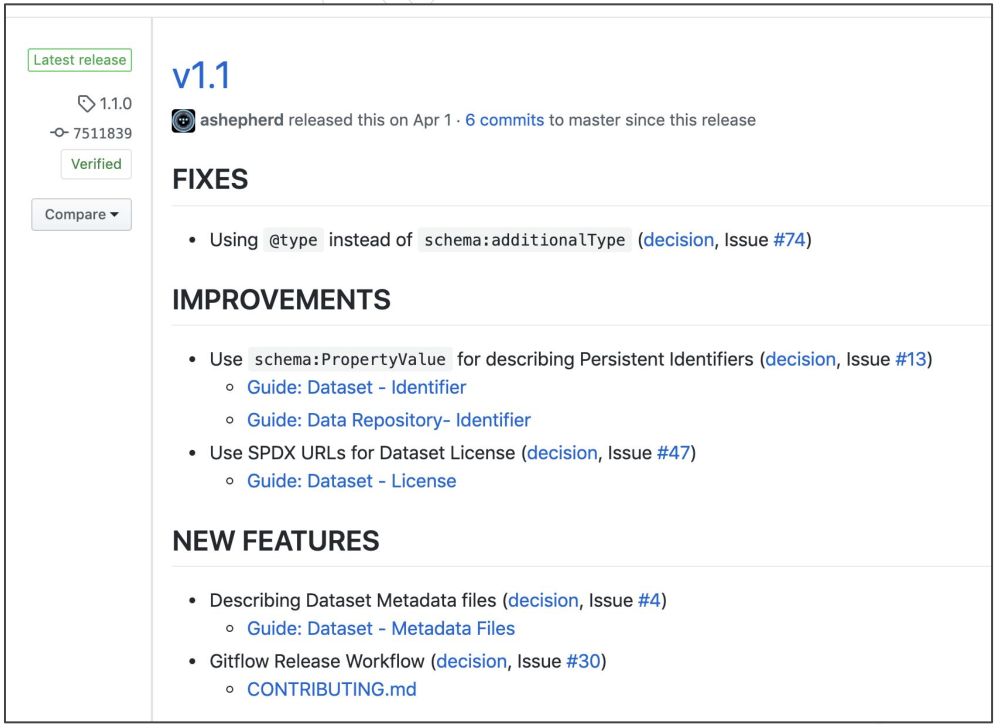
DOI 10.5281/zenodo.2628755

github.com/ESIPFed/science-on-schema.org/releases/tag/1.1.0

- **Persistent Identifiers**
- **Dataset Licenses**
- Distinguish Dataset **Metadata** from **Data files**

7 contributors:

Matt Jones
Dave Vieglais
Stephen Richard
Ruth Duerr
Lewis John McGibbney
Charles Vardeman II
Douglas Fils
Adam Shepherd



v1.1

ashepherd released this on Apr 1 · 6 commits to master since this release

FIXES

- Using `@type` instead of `schema:additionalType` (decision, Issue #74)

IMPROVEMENTS

- Use `schema:PropertyValue` for describing Persistent Identifiers (decision, Issue #13)
  - Guide: Dataset - Identifier
  - Guide: Data Repository- Identifier
- Use SPDX URLs for Dataset License (decision, Issue #47)
  - Guide: Dataset - License

NEW FEATURES

- Describing Dataset Metadata files (decision, Issue #4)
  - Guide: Dataset - Metadata Files
- Gitflow Release Workflow (decision, Issue #30)
  - CONTRIBUTING.md

ESIP Schema.org Cluster

1. Monitor monthly updates to schema.org
2. Accept contributions and issues at Github
3. Maintain & publish updates to guidelines

@ **science-on-schema.org**

Telecons:

- 1st Monday at 5pm ET
- 4th Thursday at 2:30pm ET

WORLD DATA SYSTEM

Resources to learn more:

- How to Use RDA Crosswalks and Crosswalk Visualizations
  - To be published March 2021
- Schema.org for Research Data Managers: A Primer
  - Possibly a lightning version
  - To be published March 2021 to CODATA

These documents all together should enhance guidance documents for the actual SDO markup

WORLD DATA SYSTEM

Thank you for your time and to the Portage Network!

How to Contact Us:

Chantelle Verhey,
International Technology Office-World Data systems
Ito-ra2@oceannetworks.ca

Adam Shepard,
BCO-DMO
ashepherd@whoi.edu

January 20th, 2021, wds-ito.org

WORLD
DATA SYSTEM