International Journal of Big Data Management

ISSN online: 2631-8687 - ISSN print: 2631-8679
https://www.inderscience.com/ijbdm

Schema.org for research data managers: a primer

Karen Payne, Chantelle Verhey

DOI: 10.1504/IJBDM.2022.10048569

Article History:

Received:	31 March 2021
Accepted:	21 May 2021
Published online:	23 January 2023

Schema.org for research data managers: a primer

Karen Payne and Chantelle Verhey*

International Science Council,
World Data System International Technology Office,
#100-2474 Arbutus Rd., Victoria, BC V8N-1V8, Canada
Email: ito-director@oceannetworks.ca
Email: cverhey@oceannetworks.ca
*Corresponding author

Abstract: Data managers are currently investigating the value proposition of Schema.org (SDO) with a lightweight implementation that couples the SDO vocabulary with indexing Google Dataset Search. This pathway essentially 'webifies' dataset search and syndication and enhances dataset discovery. The Primer was created to layout the SDO landscape, where it came from, what is driving its uptake in research data management, and how it works in broad strokes. It was designed to introduce individuals with little technical knowledge to the benefits and importance of Schema.org and aimed to be a 'one-step-back' from the numerous guidance documents that are being produced in the RDM community. This document describes the mark-up process in very simple terms, provides current methods of adoption by the research community, describes related technical areas relevant to data managers and lists what organisations they should follow to keep apprised of this work.

Keywords: Schema.org; SDO; interoperability; Google Dataset Search; GDSS; indexing; mark-up; research data managers; extensions; international technology office; semantics; ontologies; Semantic Web.

Reference to this paper should be made as follows: Payne, K. and Verhey, C. (2022) 'Schema.org for research data managers: a primer', *Int. J. Big Data Management*, Vol. 2, No. 2, pp.95–116.

Biographical notes: Karen Payne is an Associate Director for International Technology for the World Data System, a component of the International Science Council. In this role, she supports WDS repositories as they develop data services, and coordinates member contributions to the global open science commons. Prior to joining the University of Victoria, she worked at University of Georgia providing data services to the humanitarian community involved in disaster relief and recovery activities. She obtained a joint PhD in Geography and Engineering from the Australian National University investigating artificial intelligence techniques for classifying satellite images.

Chantelle Verhey is a Research Associate for the World Data System – International Technology Office hosted at Ocean Networks Canada. She holds a Masters of Science in Environmental Management from the University of Reading in the UK, and was dedicated to researching forest fire trends in the Canadian Boreal Forest. After her research was completed, she moved on to work at the University of Waterloo as the Data Manager for the Polar Data Catalogue. Currently, she is combining her research and work experience to investigate the usefulness of Schema.org implementation as it attempts to open the discovery of datasets to a much broader audience.

1 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, documented and widely available to the research community and the public more broadly. This mandate is encapsulated in the FAIR principles for data, and the TRUST principles for data repositories (Guha et al., 2016; Lin et al., 2020). FAIR principles promote activities that ensure data is findable, accessible, interoperable and reusable, while adhering to the TRUST principles of transparency, responsibility, user focus, sustainability and technology (Guha et al., 2016; Wilkinson et al., 2016; Lin et al., 2020). Semantic mark-up is a core technology we have at our disposal to ensure that data managers adhere to these principles. However, data managers who are new to semantic mark-up may find the landscape baffling. This paper is a primer targeted to data managers who need a simple, straightforward and accessible introduction to semantic mark-up, beginning with Schema.org (pronounced 'schema dot org' or SDO). While we will talk about semantics generally, our focus is on why there is so much discussion about SDO in the research data management community. We will describe where it came from, what is driving its uptake and how it works in broad strokes. By the end of the paper the reader should be able to read and understand more technical guidance documents, and learn more about what organisations they should follow to keep apprised of this work.

2 History

The idea of the Semantic Web long predates the arrival of SDO. Berners-Lee and Fischetti (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. Importantly, it defines not just links between content (like hyperlinks) but also the nature of the connections and the meaning of the linked content. As data managers and scientists with increasing volumes and variety of data to manage, discover and analyse, it is imperative that we adopt semantic mark-up and related technologies so that we can automate our workflows. We will describe how semantics work in lay terms, and give more details about SDO below, but in short, SDO is a controlled vocabulary used to mark up content in a webpage in a way that search engines can understand. The core SDO vocabulary is under development and growing. It can be used to describe a lot of different things, called 'object types', that are referred to in a webpage, including: recipes, job posts, reviews and in our case datasets and data catalogues.

SDO is a specific example of the broader class of semantic technologies. SDO was created by four major search engines in 2011: Google, Microsoft Bing, Yahoo and Yandex (Guha, 2011). Sometimes referred to as a library, SDO is a set of terms that describe common things or objects, especially common objects people search for on the web. The four search engines all agreed to use the same terms, in the same way, to refer to the same objects. The SDO mark-up library is understood by all four search engines.

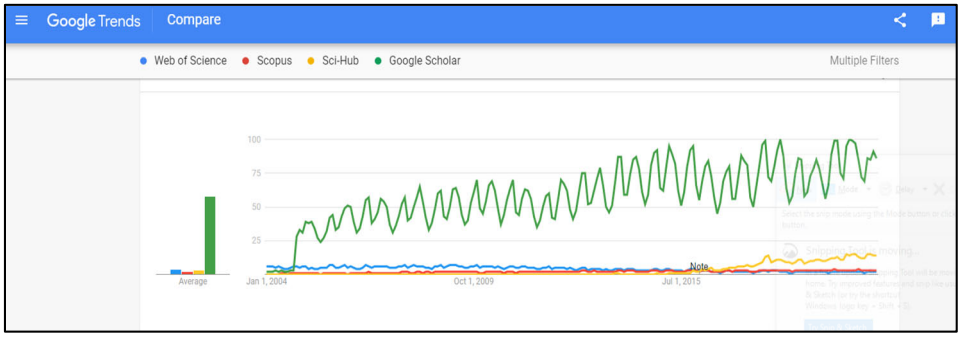
SDO maintenance and development is conducted by two groups: a steering group and a community group. The steering group is responsible for the general high-level oversight of SDO, while the community group is tasked with updating and preparing new releases. As of this writing, SDO has released version 11.0, and similar to previous releases it

consists largely of minor bug fixes identified by community members. There are times when releases contain much more extensive enhancements, such as the set of releases beginning with version 7 in March 2020 that extended the vocabulary to support identifying COVID-19 resources like special announcements about disease prevention and the location of COVID testing facilities (SDO, 2020). The *SDO community group* conversations and content is hosted by W3C (but is not a W3C body) and anyone can ask to join. The steering group participants are representatives from the four search engines.

3 The primary driver: Google Dataset Search

Google Dataset Search (GDSS – the acronym is used widely within the RDM community) 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. As shown in Figure 1, trends in searches for GS have far outstripped searches for other popular research publication repositories Web of Science, Scopus and Sci-Hub since 2004.

Figure 1 *Trends* in searches for GS compared to Web of Science, Scopus and Sci-Hub since 2004 (see online version for colours)



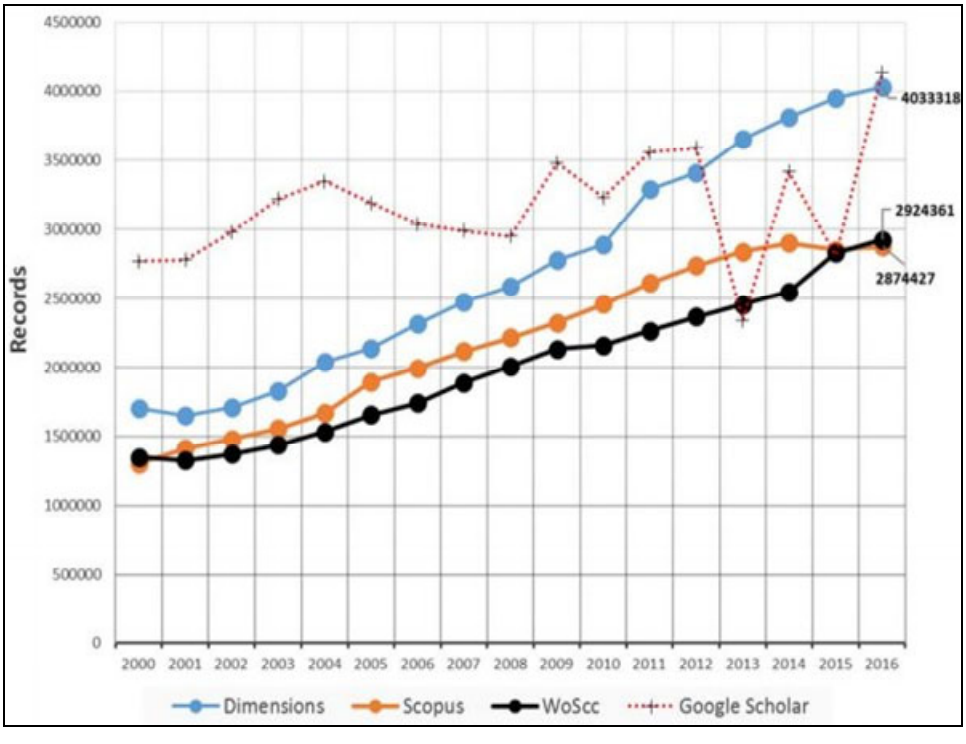
Notes: Using Google Analytics this comparative analysis displays searches for GS, Web of Science, Scopus and Science Hub in a time series (*source*).

It is worth noting that the start date for *Google Trends* and GS was 2004, also coinciding with the start date of *Scopus*. This affects the available data for these comparisons prior to 2004. Furthermore, the creation of *Sci-Hub*, which was in 2011, does alter the metrics as it still is a relatively new platform. As of this document's creation, a statistical analysis of the number of searches per year within these search engines versus the number of searches for the platforms was not conducted. This is due to the lack of reporting of these types of engagement metrics from these platforms. Figure 2 from Orduña-Malea and Delgado-López-Cózar (2018) shows that the number of publication records indexed by GS outstripped Dimensions, Web of Science and Scopus.

Search engine optimisation (SEO) is a long-standing trend in web development. Data managers who use SEO technologies to make their data more widely available are recognising larger trends in the publishing community more broadly. Within the USA, from January 2016 to the end of June 2020, employment in internet publishing and web

search portals increased by 48%. Over the same period, employment decreased 38% in newspaper publishers, 25% in periodical publishers, and 17% in book publishers (Bureau of Labor Statistics, US Department of Labor, 2020).

Figure 2 The number of publication records indexed by GS, Dimensions, Web of Science and Scopus (see online version for colours)



Source: Orduña-Malea and Delgado-López-Cózar (2018)

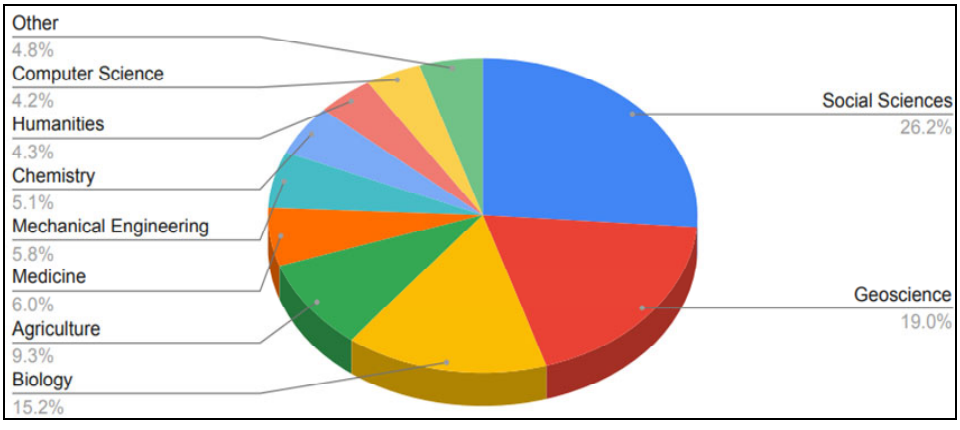
The natural appeal of both GDSS and GS stems from the user-friendly interface that *Google* creates, making it an easy go-to for finding content (Gusenbauer, 2019). It should also be noted that in addition to ease of use and accessibility, a determining factor when choosing a search engine is cost. For example, both *Google* and *Sci-Hub* are free to use, while *Web of Science* and *Scopus* are subscription-based resources, which can drive usage metrics (Martín-Martín et al., 2020). Furthermore, 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.

4 Adoption by the research community

GDSS launched in beta in September 2018. Analysis of GDSS holdings indicated that between inception and late March of 2020, the number of datasets grew from 6 M to 28 M, with the vast majority of datasets described in English (Benjelloun et al., 2020). The two entities providing the largest number of datasets are the commercial economic

data provider ceicdata.com (~20%–3.7 million) and the US Federal data group at data.gov (~17%–3.1 million). As shown in Figure 3, the largest uptake has been by geosciences and social sciences.

Figure 3 Distribution of datasets by broad coverage topic, inferred from dataset metadata and the web page itself (see online version for colours)



Source: Benjelloun et al. (2020)

4.1 Extensions

As of this writing, the *SDO* vocabulary consists of 778 types, 1,383 properties, 15 data types, 73 enumerations and 367 enumeration members (Welcome to Schema.org, 2020a). Only a small fraction of the vocabulary is used for metadata mark-up. In reality, for data managers, *SDO* is lightweight and lacks any meaningful detail about the datasets it describes. This situation is a double-edged sword. While the general layer of SDO does not contain domain knowledge, it is fairly easy to implement and holds the promise of wider discovery of their data via GDSS. For repositories that have the capacity and motivation to go further, there are options for domain-rich SDO extensions developed by different research communities.

Earlier in the development of SDO, domain-specific communities were allowed to develop and submit *extensions* to SDO as candidates to be included in the core releases. The *biomedical community* submitted an extension to SDO and to date is the only domain-specific extension designed to support metadata that has been integrated into the SDO core. This extension includes the most generic types of entities related to health and the practice of medicine; there are far more detailed semantic resources for health sciences as we discuss below. This core extension was developed through expert reviewers from institutions such as Harvard, Duke and several health websites, in collaboration with the W3C healthcare and life sciences communities to help bridge the complex worlds of web structures and medicine/healthcare [documentation for health/medical types (SDO, 2020)]. SDO is not accepting requests for extensions at this time, but reserves the right to adopt extensions that are externally developed and maintained by community groups in the future. Since the roll-out of SDO, a few domain communities have created extended domain-specific vocabularies to enhance discoverability and topic-related searches.

In addition to the biomedical extension that has been included in the SDO core, a related extension was created for the life science community by *bioschemas.org*. Bioschemas aims to improve the findability on the web of life sciences resources such as datasets, software and training materials. It utilises the base SDO, and encourages community members to be consistent users of the mark-up. Use of bioschemas has been endorsed by the European Research Council in their open research data and data management plans policy (European Commission, 2019).

For the earth science community, much work has been done developing the conveniently named *science-on-schema* extension (Jones et al., 2021). This extension was spearheaded by the *Earth Science Information Partners* (*ESIP*) community, and maintained by the ESIP SDO cluster group. The Geoschemas.org (2020) community utilises the science-on-schema base, but has also developed a temporal semantic SDO set, and initially began as an NSF *EarthCube* initiative (Peckham and Sheehan, 2017). ESIP has published a guidance doc for implementing their science-on-schema extension (Jones et al., 2021).

In addition to extensions, numerous groups, both generalist and domain-specific, have produced guidelines, tools and resources that support data managers interested in SDO mark-up. For example, the Polar Data Discovery Enhancement Research (*POLDER*) has published *documentation* and examples for SDO mark-up specifically for the Polar community. The US National Institute of Standards and Technology has collected *a list of projects and platforms* used by researchers that support SDO (Trust, 2018). The general purpose data repository software *CKAN* has been extended to include SDO mark-up for DCAT metadata. SDO mark-up is not exclusive to dataset metadata (Oderbolz, 2018). The CodeMeta Project (2020) also promotes the use of SDO mark-up for metadata that describes scientific software. The Brokered Alignment of Long-tailed Observations (*BALTO*) project funded by EarthCube has created an extension known as *Hyrax*, which takes in datasets from different providers in the earth sciences and publishes a dataset landing page with SDO mark-up in JSON-LD, ready to be parsed by GDSS (Doughty, 2020; Peckham and Sheehan, 2017; Hyrax, 2020). Finally, it is also worth noting that automating the addition of semantic mark-up to existing metadata is also an active area of research. For example, the *CGIAR* Platform for Big Data in Agriculture is currently looking into text mining tools to parse metadata in the agricultural community and add relevant mark-up as part of the development of their Global Agricultural Research Data Innovation Acceleration Network (*GARDIAN*) platform that includes data, publications and tools for managing data (Big Data, 2020). Multiple working and interest groups within the Research Data Alliance are examining semantic schemas, interoperability and controlled vocabularies, and are a good place to learn more about these initiatives and more.

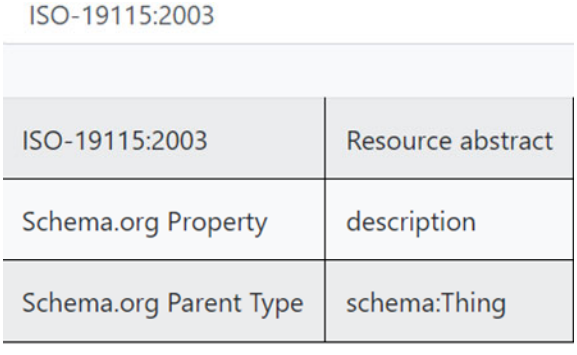
5 How does it work?

Generally speaking, the workflow is this: repositories have a database of metadata that describe their holdings. Data managers can use that set of metadata to generate a series of landing pages, one per dataset. That landing page is a web page generated from the database of metadata that describes the dataset and includes either a pointer to the data resource that can be found at the repository or instructions on how to access the data. The

landing pages are marked up with SDO terms. Bots then parse the web pages, index them and include them in GDSS. Importantly for those who manage sensitive or restricted data that 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.

Standard metadata formats like DCAT, ISO-19115, FGDC, CSDGM and others define descriptive metadata properties including title, subject, genre, author and creation date. In order to mark up the landing page with SDO terms, the data manager has to decide what metadata properties (the terms used in standard metadata formats) are equivalent to what SDO terms. Finding equivalent terms across standards is known as crosswalking. Figure 4 displays a crosswalk between the well-known metadata standard ISO-19115 'resource abstract' (from gmd:MD_DataIdentification/gmd:abstract) to the *SDO parent type* 'thing' (schema:Thing/schema:description). The gmd:MD_DataIdentification and schema:Thing are the entities that carry or are refined by the gmd:abstract and schema:description properties.

Figure 4 A snapshot from the RDA crosswalk visualisations filter table indicating that the ISO term 'resource abstract' is equivalent to the SDO property 'description' which refers to the SDO type 'thing'



Source: Welcome to Schema.org (2020b)

To support data managers interested in SDO adoption, and to promote consistent use of terms across research domains, the RDA Research Metadata Schema WG has collected approximately *15 crosswalks* between SDO terms and metadata properties used by data managers. The WDS-ITO has created user-friendly *visualisations* of the collected crosswalks to promote consistent community implementations and to facilitate the derivation of new crosswalks. The EarthCube Hyrax extension supports inputs from multiple metadata standards and members of the RDA Research Metadata Schema WG are currently investigating the EarthCube crosswalks to compare with the set collected within RDA (Hyrax, 2020; Peckham and Sheehan, 2017).

When learning about semantic mark-up, it is helpful to contrast it with other types of mark-up. You are likely familiar with HTML mark-up. HTML mark-up instructs your web browser how to display content. For example,

```
<!DOCTYPE html>
<html>
<body>
<h1>New Whale Species!</h1>
<p>A new <mark>whale species</mark> was discovered off the coast of <b>Mexico</b> today.</p>
</body>
</html>
```

will display in your browser in Figure 5.

Figure 5 The user friendly content of the HTML mark-up above (see online version for colours)

Figure 6 Sample result of Google Search for the word 'table' (see online version for colours)

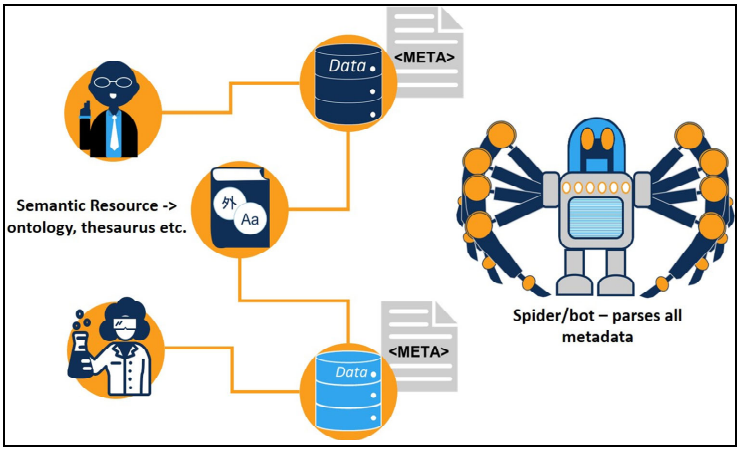
Tags surround content. <*tag*> indicates the start of mark-up and a tag preceded with a 'slash' (</*tag*>) indicates where the mark-up ends. HTML mark-up works in conjunction with cascading style sheets (CSS) to identify sections or types of text and render it in a browser. In our example, the <h1> tag indicates the title, which is displayed in large letters; the <mark> tag indicates what should be highlighted and the bold (<b>) tag

indicates what should be displayed in bold letters. HTML and CSS mark-up is intended for humans reading content and is mostly about visual rendering. In contrast, semantic mark-up is intended for machines and is intended to attach more meaning to content.

Imagine you do a simple Google Search for the word 'table'. Your search results will of course include furniture, but it will also include results about tabular structures to organise information like spreadsheets and charts. For illustration purposes, the authors did a search for 'table' and received additional information about local events and businesses and something about cricket that is outside of our knowledge area.

Google does what it can to find search results that are relevant to you. For example, it can use your search history to push things you search for regularly to the top of the results. It can geocode your IP address to find things near you. Users will also find prompts at the bottom of the page asking for context words to narrow the search ('searches related to table: kitchen table, table math'). Semantic searches remove the need for ancillary information and take the guesswork out of this (and other) processes.

Figure 7 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) (see online version for colours)



Notes: If someone else publishes additional content that points to the same online semantic resource, then a machine can automatically identify those two publications, with certainty, as being related. 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.

The Semantic Web relies on the existence of online definitions. These semantic resources can be a simple list of controlled terms, a vocabulary or more complex ontologies. It relies on people who publish content on the web to mark up or enhance their publications to say "this piece of content is about this subject, which is defined by this online dictionary." To refer back to our example using search as the primary use case, if I was only interested in finding web pages that contained tables (or images, videos, breadcrumbs, licenses or other types of page elements), then I could instruct a semantic search engine to: "please find all tables as defined by https://schema.org/Table" and it

would only return web pages that have marked up and identified elements as tables by pointing to SDO, and I would never receive results about furniture or cricket.

Again, SDO is an evolving vocabulary used to mark up content in a webpage so that it can be understood by all search engines. It describes any number of things, or 'object types' – for this primer we will focus on datasets. Data managers can utilise SDO terms to describe datasets by marking up metadata elements. A data manager can add terms from the SDO vocabulary to a webpage, in our case a dataset landing page or metadata document, and define values for different metadata elements – like dates, titles, data provider, abstracts, location – whose meaning is defined by the SDO vocabulary. Google spiders then collect and index all of that information from datasets described at repositories all over the world. The index is exposed in GDSS so that users can discover it.

There are only two terms that are required as part of a *dataset* description to be included in the GDSS index: *name* and *description*. This is the title of your dataset and a brief abstract. There are *additional terms that Google recommends* you include as part of your metadata mark-up, including but not limited to the dataset creator, download information, unique identifier, license and variables measured (Google Search Central, 2020b).

SDO supports three different mechanisms or formats to add vocabulary terms to a webpage: microdata, RDFa and JSON-LD (W3C, 2015). Both microdata and RDFa are a set of tags and HTML5 extensions that are embedded inline in HTML webpage elements. In the same way that the html tag<h1> instructs a webpage what the title of the document is, RDFa and microdata tags associate an SDO metadata element with a text value:

```
<div itemscope itemtype="http://schema.org/Dataset">
        <span itemprop="name">
        <b>Ocean Networks Canada: Expedition 2016 Wiring The Abyss</b>
        </span>
    </div>
```

The above mark-up instructs a bot that the item property 'name' refers to the name of an item of type dataset (defined at SDO). In this case, itemprop=name is the SDO element, 'Ocean…The Abyss' is the value for that element.

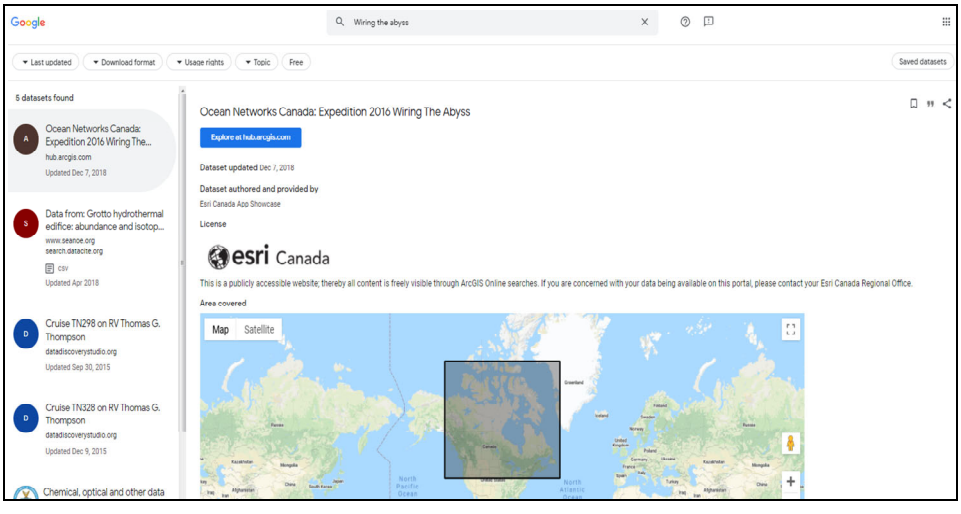
Unlike the microdata and RDFa serialisations (formats), JSON-LD is the format preferred by Google and is contained in a script block at the head of the page (Sefton et al., 2020; W3C, 2015). This is a sample of the same dataset described, with a few additional SDO terms, in JSON-LD, followed by the same dataset displayed in GDSS:

```
<script type="application/ld+json">
  {
    "@context": "https://schema.org",
    "@type": "Dataset",
        "name": "Ocean Networks Canada: Expedition 2016 Wiring The Abyss",
        "description": "Expedition 2016 Wiring the abyss. Join the Expedition!! Explore
        the ocean depths and engage with scientists and explorers in real time. It's your
        turn to experience the mystery, power and beauty of the deep sea. [Ocean
        Networks Canada]",
```

```
        "datePublished": "2018-12-07",
        "license": "https://apphub-esrica-apps.opendata.arcgis.com/datasets/
        1cb86bc3ea234b20a5a91b0c64e25447#",
        "inLanguage": "English",
        "creator": {
            "@type": "Organisation",
            "name": "Esri Canada App Showcase"
                    }
    }
</script>
```

In order to have your marked up metadata landing on pages indexed by GDSS, your site is best served by including a sitemap. A sitemap is a file that provides information about your webpages; it can be described as the equivalent of your site's outline (Marie, 2018). A sitemap is a text document (generally xml or html with mark-up) that lists the pages that are contained within a website domain. Sitemaps help the Google bots crawl over your pages in an intelligent way, acting as a sort of traffic cop and pointing out important features.

Figure 8 A view of the sample data in GDSS (see online version for colours)



Sitemaps make it easier for web crawlers to access the information on your site, and give the web developer an opportunity to inform crawlers what is most important on your site and the relationship between the information presented (Google Search Central, 2020c). Additionally, important information such as languages, last updates, and versioning of the site is typically included in sitemaps and available for crawlers to parse. Although it is noted that even if a sitemap is implemented, the crawler may not be able to pick up all the information presented. Rather, a sitemap can improve the crawling of larger or more complex sites, or more specialised files, and is strongly encouraged by Google.

Google has released *documentation* outlining the process of implementing SDO and recommendations for dataset mark-up (Google Search Central, 2020b). They also provide

online tools to test mark-up before deploying, and to test URLs after deployment. Specifically, Google's popular *Structured Data testing tool*, once slated for deprecation, has been kept but is scheduled to be moved to the SDO site by April 2021 (Levering, 2020). The purpose of the Structured Data testing tool is to check syntax and compliance with SDO standards. More recently, in July of 2020, Google released the *Rich Results test tool* (Samet, 2020). The Rich Results tool will be maintained on Google's own site and is designed to show users Google Search rich result types, discussed further below. In addition to Google's documentation and tools, we also recommend the forthcoming guidance documents in development by the Research Data Alliance *Research Metadata Schemas* Working Group for more information about how to utilise SDO to describe datasets.

6 Other driving factors and use cases for data managers

6.1 Aligning semantic mark-up across domains

For data managers, SDO serves at least three functions. First, as we have discussed, it immediately opens our data holdings to a wider audience via GDSS. Second, for data managers who are new to semantics, it is a high level, simple introduction to semantic mark-up and serves as a gateway to more complex ontologies and workflows. Finally, it is a good starting point for aligning ontologies (and by extension our datasets) across domains. It is the first step towards a common language for data on the web. It is this last function that the *science-on-schema*, *geoschemas* and *bioschemas* projects are hoping to leverage. While Google and other search engines have backed SDO, it is just the tip of the iceberg. SDO is a very, very lightweight ontology and it is left to the domain-specific research communities to develop more expressive ontologies that adequately describe their datasets. 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*, which is translated into four languages (WMO, 1959; Stocker, 2017).

The creation of controlled domain-specific vocabularies is fairly labour intensive; it requires quite a bit of collaboration within scientific communities to reach a consensus on which controlled vocabularies should be included in the extension, and how they should be implemented (Jonquet 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 may alleviate a lot of work. It is important to recognise large initiatives like the robust set of PaST (2020) from the Paleoclimatology community and the Medical Subject Headings (*MeSH*) from the National Library of Medicine (2020), which are important pillars in scientific semantics. No matter what the domain, it is best practice to re-use vocabularies before you build your own. If you are looking for vocabulary servers, we recommend BARTOC (2020), where you can search for both vocabularies and semantic registries. Also, FAIRsharing (2020), an

RDA-endorsed repository, is a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies. To further enhance established ontologies, mappings can be developed between prominent ontologies that can serve larger communities and support interdisciplinary research (Laadhar et al., 2020). For example, if a repository can easily detect term reuse between ontologies, these 'overlaps' are very useful as they can be used by developers to identify and create formal mappings.

6.2 Syndication

As we have seen, repositories are interested in using SDO to make web content describing datasets more discoverable. Both search engines and open-source tools have used it successfully to build an open ecosystem for various types of content (Guha et al., 2016). SDO allows for 'webification' of dataset landing pages as one pathway to publishing. Traditionally, data managers have used harvestable metadata services like OAI-PMH and CSW to achieve the same thing (Open Archives Initiative Object Reuse and Exchange ORE, 2008). Metadata aggregators like OpenAIRE and Canada's Federated Research Data Repository (FRDR) can read those protocols, parse a list of metadata records and copy them to a central store that is available to be searched. The harvesting process happens on a schedule, say once an evening, and includes the ability to compare metadata already harvested with newly published metadata that needs to be harvested. That way the aggregator is not copying the same metadata records over and over again.

Some data managers are investigating replacing their older harvestable metadata services with SDO mark-up and associated tools to enable harvesting across federated repositories. For example, DataONE provides a unified search portal over its member repositories' holdings (many with their own search portals) by indexing the core SDO terms used in metadata repositories across DataONE and opening them up for search while providing a much larger set of domain-specific metadata tags for advanced search options (Mecum et al., 2018b). Similarly, in the same vein as GDSS, the *GeoCODES* project (formerly known as NSF EarthCube's Project 418) has created a search engine to crawl metadata that has science-on-schema mark-up (which includes SDO terms) and create a searchable index of data for the purposes of data discovery (Mecum et al., 2018a). The advantage of this is that researchers can search for terms that are relevant to domain-specific studies, in addition to exposing their data to be indexed by GDSS (Potter et al., 2020).

We noted above that Google recommends a suite of terms be included in the mark-up for a dataset landing page to be included in their index of searchable datasets. One of these recommended terms is *variableMeasured*, which indicates what observable properties are reported in the dataset (temperature, leaf area index, etc.). The permitted value of that property is anything that can be typed into an open text field. In other words, it does not refer to a controlled vocabulary of standard observation types (i.e., free text). If someone marks up their metadata with 'temperature' (or 'temp' or '°C') as a variableMeasured, the consumer (either human or machine) does not know if that is air temperature or sea temperature; they do not know what instrument was used to measure; they do not know if it is an average of measurements, or a single measurement. All of these questions are currently being investigated by the *i-ADOPT RDA Group* who are looking at how to describe measured parameters. Initially focusing only on environmental

applications, they are aligning approximately 100 relevant vocabularies and have also aligned their recommendations with the VSSIG Semantics repository group and the *FAIRsFAIR semantics criteria* (FAIRsFAIR, 2019). Similarly, the *Canadian Consortium for Arctic Data Interoperability* are also investigating semantics for measured variables as part of their work to promote polar research. Within the ocean sciences community, measured variables such as *temperature*, *pressure*, *turbidity*, *currents*, etc. (Cameron et al., 2009; Stocker, 2017) are imperative for proper semantic distinction in datasets. In the future, data managers will need to decide if they want to use these types of controlled vocabularies alongside SDO terms like *measurementTechnique*, or if they are better served with more complex ontologies developed in their domain. Adding this additional layer of semantics into our metadata mark-up will be important for machine processing data, as described below.

6.3 Other types of ontologies

This document has focused on the use of SDO for dataset discovery; however, there are many other activities that repository managers are engaged with that will benefit from integrating more complex semantic resources to make use of intelligent agents and automated processes. What follows is a short list of what you may want to be aware of for the future.

6.3.1 Fair Digital Objects

Developers of FDOs (2020) seek to "bind all critical information about an entity in one place and create a new kind of actionable, meaningful and technology independent object." In practice, a digital object is represented by a bit stream (what we generally think of as a dataset, but can be any digital object), and stored in a repository. A digital object can be singular or it may be part of a larger collection of digital objects. It has a globally unique persistent identifier (PID) and is described by metadata (Schwardmann, 2020). FDOs are packaged with semantic metadata that enable interactions with automated data processing systems. One example is the addition of 'data type' to the metadata description. Data types act in the same way that media types or MIME types do with browsers. A browser that encounters a piece of content that is tagged with a MIME type understands how to manage or render that content. For example, a piece of content tagged with the extension .avi is an audio file, of MIME type video/x-msvideo. Automated data processors are being built that will look at data type registries to 'understand' how to interact and process FDOs tagged in their metadata with their data types. FDOs will need more than just the data type to be understood; in addition, they will require machine actionable metadata descriptions with formal registered semantic ontologies about licenses, the location of unique identifiers, field properties and other information necessary to execute defined operations that can be performed on them in the absence of human intervention. Data managers are particularly interested in FDOs as a way to store and interact with data assets. However, FDOs can refer to all kinds of digital information such as software, configuration files, representations of persons, institutions, semantic concepts, etc. (De Smedt et al., 2020).

6.3.2 Data visitation

FDOs are a core component of what some researchers are referring to as Fair Data Points (FDPs). These are locations on the web where algorithms can 'visit' data and process data (FDOs) in situ. FDPs are data repositories with 'docking' capabilities for virtual machines (VMs) that come to 'visit' and query the data locally, overcoming the overhead of computational loads required when downloading or copying data, and avoiding legal issues around moving data between jurisdictions (Go Fair, 2020). From a data manager's perspective, you can think of this as our need to manage metadata related to data, and metadata needed to manage executables, scripts and tools to process that data, all of which require semantic infrastructure. Some data visitation structures begin with a harvestable metadata service, and build visitation structures on top to allow for querying and processing.

6.3.3 Workflows

Whether processing remotely in visitation, or processing local copies, there is a need to document steps taken when researchers analyse data. This has promoted a lot of work on systems that can create reproducible workflows or papers. Scientific infrastructure developers have a vision of all digital science assets being freely available, interoperable and reusable online. This vision is a movement away from static peer-reviewed publications and a move towards publications as a re-runnable application. Moreover, the application should have plug and play components, so that a researcher can run an analysis over one set of data, and another scientist can run the same analysis technique over a different dataset, while a third reproduces the work of the first researcher, but tweaks the parameters associated with the analysis to see the result. To achieve this vision, additional semantics are needed to capture computational workflows. The virtual research environment (VRE) community has identified over 280 workflow management systems available to them. The most popular VREs that are being developed for science seem to be converging on adding support for Common Workflow Language (*CWL*). CWL is an open standard designed to describe "analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments" (Amstutz et al., 2016). An example implementation can be found in the computational chemistry platform *SEAGrid*, where users can run, save and re-run analyses with reusable workflows. The workflow itself is a digital research object.

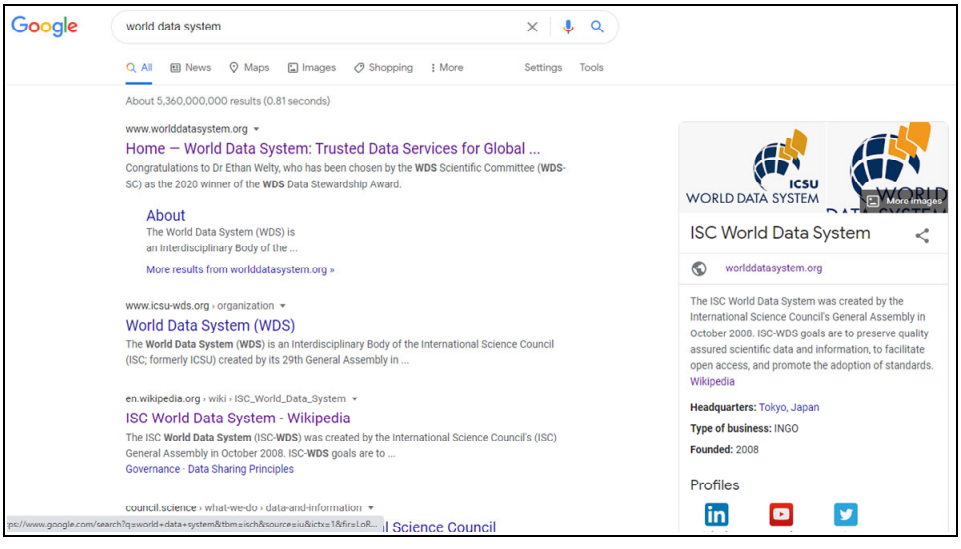
CWL also recognises the need to define the provenance of the inputs to a workflow and the workflow itself, and has worked to incorporate another W3C standard, *PROV*, to create *CWLProv*, adding support for tracking provenance of items used and modified in the workflow by recording the provenance of digital objects using linked data standards (Amstutz et al., 2016). This last enhancement is very important not only for understanding when workflows get edited, but it is also very important when working with dynamic data feeds from live sensors. Leipzig et al. (2020) provide a robust review of this work and identify metadata standards that support reproducible computations across data, tools, notebooks, pipelines and publications.

7 Future considerations for data managers

7.1 Ranking

For web developers, one of the allures of SDO is that it enables Google to return Rich Snippets (also known as 'Rich Results') to user searches. Rich Snippets are normal Google Search results with additional data displayed, and often highlighted or promoted with better and more obvious formatting. Rich snippet call outs are very user friendly and may include standard 'promotional' boxes with images, maps, short descriptions, etc. displayed in Figure 9.

Figure 9 Example of a rich snippet from the ISC World Data System to the right of a Google general search, including images and information from Wikipedia (see online version for colours)



Users are often drawn to Rich Snippets and may ignore all other search results. To date, we have not seen the use of Rich Snippets when searching for data either in the general Google Search engine, or as part of GDSS; however, it is worth keeping an eye on this in the future. There is a potential danger that data managers would be in competition to be the *de facto* data source populating content in a rich snippet, in the way that Wikipedia entries and Google Maps are *de facto* content providers for rich call out boxes generated in response to Google Search results. Also, some Rich Snippets are not very obvious about the source of the information in the callout. There is a danger that the contributions from scientific researchers and their institutions get lost, while Google controls access to what it deems as relevant and important. It would be problematic if scientists or data managers were put in a position of spending time rating datasets with gold stars in order to inform Google what is important, or if that function was relegated to non-scientists to determine which datasets have the most value. The importance of data should be based on its purpose and content and not on how readily it was packaged for easy consumption within the Google environment. Rich Snippets aside, it is unclear how Google chooses

the order of results and how the datasets are ranked – what dataset is listed first, second, etc. – in GDSS.

SDO and Rich Snippets can function at different levels of detail. According to Monk (2018), Google has been developing a rich snippet version for open data that does not just point the user to the original data source, but instead reaches into the dataset and displays the data records or observations themselves. These open data sources would be displayed as data visualisations contained in Rich Snippets in the form of charts, graphs and tables in the search engine results pages (SERPs). This would allow users to see datasets in call out boxes on pages in the SERPs, rather than having to navigate the original source of the data. This has the potential to cause quality assurance concerns. False, misleading or erroneous data published from an untrustworthy source using SEO structures could dominate the top of search results. Google is a proprietary organisation and is not transparent about how their algorithms work. In contrast, while the open data movement promotes open and trustworthy repositories and encourages developers to access and visualise their data stores, this type of implementation poses a risk to scientific discourse. This could create an avenue for bad actors to infiltrate and degrade the important work of research data managers, similar to the reports of nefarious groups hijacking submissions to peer-reviewed journals, and it is easy to imagine scammers creating data feeds to amplify climate change and vaccine deniers (Duranni, 2020).

Some data managers who have implemented SDO have remarked that the indexing process is not as robust as hoped. Some worry about the amount of time it takes to get indexed after the landing pages are marked up. One data manager did experiments and worked with the creators of GDSS directly to determine how to optimise the speed of indexing, but generally speaking it remains an obtuse process. Crawlers will search entries when they get around to it and it is impossible to know how long it will take – in some cases it has reportedly taken months. To be fair, Google has provided documentation on how to request indexing for individual or multiple pages, assuming a sitemap has been published and has deployed a *site* that allows users to request that their site be indexed (or re-indexed) (Google Search Central, 2020a). However, Google has also indicated they "prioritise the fast inclusion of high quality, useful content" and it is not clear where dataset landing pages are on their priority list. As of this writing, it appears that the Request Indexing service has been unavailable since 14 October 2020 while Google makes technical updates, likely associated with their recent Rich Snippet tools release and migration of the Structured Data testing tool described above (Google Search Console Help, 2020). We have heard similar complaints from developers who note that there seems to be too few dedicated resources to manage and investigate the issues submitted in the SDO *github* repository (Kriedler, 2020). Perhaps this is understandable given that the day-to-day operations of the SDO project are run by a volunteer community group. As the research data management community continues to invest in SDO with more repositories completing the implementation, these problems may be resolved if the steering group is convinced to increase resources to the project.

For all of the reasons described above, we believe it is important that the scientific community do more to be seen as a constituency of the W3C and wider developer community. We need to do a better job of reaching out and including our work in use cases documented and read by web developers. For instance, as part of the RDA working group on research metadata schemas, we propose that examples of SDO mark-up of research metadata be included in W3C community documentation. There are multiple W3C Community Groups that are focussed partially or entirely on SDO improvements

for domains. Each group has their own way of working, and coordinate via the main SDO community group (https://schema.org/docs/about.html#cgsg), and would be well served by increased representation from the RDM community. As Le Franc et al. (2020) have recommended, a common governance model for semantic artefacts across scientific communities would be helpful.

7.2 The role of harvesting

Traditionally, data managers have relied on harvesting protocols to syndicate their metadata. To date, we have not seen an investigation comparing the relative effectiveness and impact of SDO to harvestable services on data uptake. Historically, harvestable metadata services have been the go-to technology for ensuring interoperability between repositories. Ultimately, both are viable options for ensuring the data hosted is available, and some repositories have elected to use both harvesting metadata services and SDO mark-up in landing pages to ensure that all their bases are covered. Realistically, neither option produces perfect results, so a combination approach may be the most viable, depending, of course, on the resources and skill sets available. It is unclear if we will migrate away from harvesting in favour of SDO in the future. We may be in a period where use of SDO is just not well enough understood by under-resourced data managers to take full advantage of the resources. For example, the SDO types of *DataFeed* and *DataFeedItem* have not shown up in our review of existing research metadata crosswalks. Both terms seem like a natural fit for repositories that provide metadata describing live data services via APIs, *WFS*, *WCS* or similar services (Government of Canada, 2015; Mink, 2015).

7.3 Continuing education

Once data managers use SDO to get acquainted with using semantic technologies, they will constantly have to improve their knowledge and use different technologies as engaged with semantic artefacts of increasing complexity. This creates a very specific need for ongoing capacity development for data managers. This table from Le Franc et al. (2020) breaks down the change in standards used as the semantic artefacts become increasingly complex (Table 1).

Table 1 Common technologies used in different types of semantic artefacts

Type of semantic artefact	*Currently used standards (serialisation formats and data models)*
List (terminologies, glossaries, vocabularies)	CSV, XML, JSON, SKOS
Hierarchical list	XML schema, RDF, SKOS
Thesaurus	RDF/RDFs, SKOS
Formal ontology	OWL, OntoUML, FOL, Modal logic

Source: Table from Le Franc et al. (2020).

As the data manager adopts increasingly complex semantic use cases, they will also need to contend with other technologies, specifically PIDs, if the repository has not adopted them yet. PIDs are required to make data optimally citable, and elevate them to become first-class citizens of the scientific landscape. In their 2020 review of indexed metadata in

GDSS, Benjelloun et al. (2020) noted that only about 11% of the datasets (or ~3M) had DOIs, and that the vast majority (about 2.3M) of those come from two sites, datacite.org and figshare.com. It is worth noting that DataCite automatically generates SDO mark-up for datasets that use DataCite to generate DOIs. However, the SDO entry generated by DataCite is fairly minimal because it is based on the seven required metadata fields required to generate a DOI.

Only a tiny fraction, 0.45%, of the datasets has compact identifiers (a mix of a namespace prefix and a locally unique identifier). While the combination of PIDs and semantic structures is essential for machine actionable research, as we described above, it is also a boost to the search industry. Google's intent is to use the uniquely identified, indexed metadata in combination with structured mark-up from GS to continue to build out their own knowledge graph. This will be in direct competition with similar initiatives in large publishers and community groups like OpenAIRE, and will be another potential avenue for Google to monetise the work done by the research community. To address this, one of the responsibilities of data managers moving forward will be to engage in conversations with multinational tech companies, and advocate with lawmakers about creating structures for commercial entities to support academic and scientific research from which they profit.

Acknowledgements

This study is funded under the Canadian New Digital Research Infrastructure Organization (NDRIO) collaborative agreement number 51387-50326. We would like to thank the fantastic feedback provided by both the ESIP SDO Cluster Group and the Portage Data Management Expert Group (DMEG). Thank you for allowing us to bounce ideas off of each other for this document and for providing extremely valuable feedback to ensure we produced a robust report. Specifically, a special thank you to Kelly Stathis, Kevin Ried, Steven Richard and Chantel Risdale, who volunteered their time to make suggestions and review our document.

References

Amstutz, P., Crusoe, M., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S. and Stojanovic, L. (2016) *Common Workflow Language, v1.0*, Specification, Common Work Flow Language Working Group, DOI: 10.6084/m9.figshare.3115156.v2 [online] https://w3id.org/cwl/v1.0/.

Basic Register of Thesauri, Ontologies & Classifications (BARTOC) (2020) *About* [online] https://bartoc.org/about (accessed 10 February 2021).

Benjelloun, O., Chen, S. and Noy, N. (2020) *Google Dataset Search by the Numbers*, arXiv:2006.06894 [online] https://arxiv.org/pdf/2006.06894.pdf.

Berners-Lee, T. and Fischetti, M. (1999) *Weaving the Web*, Chapter 12, Harper, San Francisco, ISBN: 978-0-06-251587-2.

Big Data (2020) *CGIAR Platform for Big Data in Agriculture Annual Report 2019*.

Bureau of Labor Statistics, US Department of Labor (2020) *The Economics Daily, Internet Publishing Employment Up 48 Percent, First Quarter 2016 to Second Quarter 2020*, 3 December [online] https://www.bls.gov/opub/ted/2020/internet-publishing-employment-up-48-percent-first-quarter-2016-to-second-quarter-2020.htm (accessed 10 December 2020).

Cameron, M.A., Wu, J., Taylor, K., Ratcliffe, D., Squire, G. and Colton, J. (2009) *Semantic Solutions for Integration of Federated Ocean Observations*, SSN.

CodeMeta Project (2020) *Motivation* [online] https://codemeta.github.io/ (accessed 10 February 2021).

De Smedt, K., Koureas, D. and Wittenburg, P. (2020) *Fair Digital Objects for Science: From Data Pieces to Actionable Knowledge Units*, Vol. 8, p.21, Publications [online] https://doi.org/10.3390/publications8020021.

Doughty, C. (2020) *Extensions to Schema.org for Structured, Semantic & Executable Research Documents*, 20 September [online] https://stenci.la/blog/2020-09-20-register-now-for-stencila-community-call-thur-24-sept-2020/ (accessed 12 December 2020).

Duranni, J. (2020) *Imposters Hijack Journal's Peer Review Process to Publish Substandard Papers*, 18 January 2021 [online] https://www.chemistryworld.com/news/imposters-hijack-journals-peer-review-process-to-publish-substandard-papers/4013050.article (accessed 19 January 2021).

European Commission (2019) *Open Research Data and Data Management Plans Information for ERC Grantees by the ERC Scientific Council*, 3 July 2020 [online] https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf (accessed 10 December 2020).

Fair Digital Objects (FDOs) (2020) *Turning Data to Knowledge* [online] https://fairdo.org/ (accessed 15 December 2020).

FAIRsFAIR (2019) *Fostering Fair Data Principles in Europe*, 1 March [online] https://www.fairsfair.eu/fairsfair-open-consultation-fair-data-policies-and-practices (accessed 10 December 2020).

FAIRsharing (2020) *A Curated, Informative and Educational Resource on Data and Metadata Standards, Inter-Related to Databases and Data Policies* [online] https://fairsharing.org/ (accessed 10 February 2021).

Geoschemas.org (2020) *Extensions* [online] https://geoschemas.org/extensions/ (accessed 10 December 2020).

Go Fair (2020) *Declaration: Virus Outbreak Data Network (VODAN) GO FAIR Implementation Network*, 10 March [online] https://www.go-fair.org/wp-content/uploads/2020/03/VODAN-IN-Manifesto.pdf (accessed 17 December 2020).

Google Search Central (2020a) *Ask Google to Recrawl Your URLs*, 14 December [online] https://developers.google.com/search/docs/advanced/crawling/ask-google-to-recrawl (accessed 17 December 2020).

Google Search Central (2020b) *Dataset* [online] https://developers.google.com/search/docs/data-types/dataset (accessed 10 December 2020).

Google Search Central (2020c) *Learn About Sitemaps*, 4 December [online] https://developers.google.com/search/docs/advanced/sitemaps/overview (accessed 13 January 2021).

Google Search Console Help (2020) *Request (Re)indexing*, 14 October [online] https://support.google.com/webmasters/answer/9012289#request_indexing (accessed 17 December 2020).

Government of Canada (2015) *Web Feature Service (WFS)*, 25 November [online] https://www.nrcan.gc.ca/earth-sciences/geomatics/canadas-spatial-data-infrastructure/standards-policies/8934 (accessed 18 December 2020).

Guha, R. (2011) *Introducing Schema.org: Search Engines Come Together for a Richer Web*, 2 June [online] https://googleblog.blogspot.com/2011/06/introducing-schemaorg-search-engines.html (accessed 10 December 2020).

Guha, R., Brickley, D. and Macbeth, S. (2016) 'Schema.org: evolution of structured data on the web', *Communications of the ACM*, Vol. 59, No. 2, pp.44–51.

Gusenbauer, M. (2019) 'Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases', *Scientometrics*, Vol. 118, pp.177–214 [online] https://doi.org/10.1007/s11192-018-2958-5.

Hyrax (2020) *Welcome to the Hyrax-1.16.2 Release Page (Updated 27 April 2020)* [online] https://www.opendap.org/software/hyrax/1.16 (accessed 17 December 2020).

Jones, M., Richard, S., Vieglais, D., Shepherd, A., Duerr, R., Fils, D. and McGibbney, L.J. (2021) *Science-on-Schema.org v1.2.0 (Version 1.2)*, Zenodo, 8 February [online] http://doi.org/10.5281/zenodo.4477164.

Jonquet, C., Toulet, A., Dutta, B. et al. (2018) 'Harnessing the power of unified metadata in an ontology repository: the case of AgroPortal', *J. Data Semant.*, Vol. 7, pp.191–221 [online] https://doi.org/10.1007/s13740-018-0091-5.

Kriedler, A. (2020) *Thread: Google and Other Tech Giants are Happy to Have Control over the Web's Metadata Schemas, But They Let its Infrastructure Languish*, 6 August [online] https://threadreaderapp.com/thread/1291509746000855040.html (accessed 11 December 2020).

Laadhar, A., Abrahão, E. and Jonquet, C. (2020) 'Analysis of term reuse, term overlap and extracted mappings across AgroPortal semantic resources', *EKAW 2020 – 22nd International Conference on Knowledge Engineering and Knowledge Management*, Bozen-Bolzano, Italy, September, No. lirmm-02945172.

Le Franc, Y., Parland-von Essen, J., Bonino, L., Lehväslaiho, H., Coen, G. and Staiger, C. (2020) *D2.2 FAIR Semantics: First Recommendations (Version 1.0)*, FAIRsFAIR [online] https://doi.org/10.5281/zenodo.3707985.

Leipzig, J., Nüst, D., Tapley Hoyt, C., Soiland-Reyes, S., Ram, R. and Greenberg., J. (2020) *The Role of Metadata in Reproducible Computational Research*, 15 June [online] https://arxiv.org/ftp/arxiv/papers/2006/2006.08589.pdf (accessed 11 December 2020).

Levering, R. (2020) *An Update on the Structured Data Testing Tool*, 15 December [online] https://developers.google.com/search/blog/2020/12/structured-data-testing-tool-update (accessed 17 December 2020).

Lin, D., Crabtree, J., Dillo, I. et al. (2020) 'The TRUST principles for digital repositories', *SciData*, Vol. 7, p.144 [online] https://doi.org/10.1038/s41597-020-0486-7.

Marie, J. (2018) *Website Sitemaps: A Comprehensive Guide*, 8 August [online] https://slickplan.com/blog/site-mapping-a-comprehensive-guide (accessed 13 January 2021).

Martín-Martín, A., Thelwall, M., Orduna-Malea, E. et al. (2020) 'Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations', *Scientometrics* [online] https://doi.org/10.1007/s11192-020-03690-4.

Mecum, B., Fils, D. and Shepard, A. (2018a) *Schema.org: Improving Access to Data through a Standardized Language*, 13 November [online] https://www.dataone.org/webinars/schemaorg-improving-access-data-through-standardized-language/ (accessed 10 December 2020).

Mecum, B., Nenuji, R., Jones, M.B., Vieglais, D. and Schildhauer, M. (2018b) 'DataONE on the web: using Schema.org and JSON-LD to enhance data search and access', *AGU Fall Meeting Abstracts*.

Mink, J. (2015) *WCS Tools Programs*, 6 November [online] http://tdc-www.harvard.edu/software/wcstools/wcsprogs.html (accessed 10 February 2021).

Monk, M. (2018) *New to Google Search: Dataset Schema*, 9 November [online] https://www.bigleap.com/blog/new-to-google-search-dataset-schema/ (accessed 18 January 2021).

National Library of Medicine (2020) *Introduction to MeSH*, 15 December [online] https://www.nlm.nih.gov/mesh/introduction.html (accessed 10 February 2021).

Oderbolz, S. (2018) *Make Open Data Discoverable for Search Engines*, 30 April [online] https://ckan.org/2018/04/30/make-open-data-discoverable-for-search-engines/ (accessed 18 January 2021).

Open Archives Initiative Object Reuse and Exchange ORE (2008) *User Guide – Primer*, 17 October [online] https://www.openarchives.org/ore/1.0/primer (accessed 10 December 2020).

Orduña-Malea, E. and Delgado-López-Cózar, E. (2018) 'Dimensions: re-discovering the ecosystem of scientific information', *Profesional De La Información*, Vol. 27, No. 2, pp.420–431 [online] https://doi.org/10.3145/epi.2018.mar.21.

Paleoenvironmental Standard Terms (PaST) (2020) *Thesaurus* [online] https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/past-thesaurus (accessed 10 December 2020).

Peckham, S. and Sheehan, A. (2017) *Collaborative Research: EarthCube Integration – Brokered Alignment of Long-Tail Observations (BALTO)*, AGS DivAtmospheric & Geospace Sciences [online] https://sites.google.com/vt.edu/balto/ (accessed 17 December 2020).

Potter, N., Gallagher, J., Travis The Robot, Kari, U., ideaesb, Rimer, R. and Korolev, S. (2020) *OPENDAP/olfs: OLFS 1.18.7 for Hyrax 1.16.2 (Version olfs-1.18.7)*, Zenodo, 22 April [online] http://doi.org/10.5281/zenodo.3762740.

Samet, M. (2020) *The Rich Results Test is Out of Beta*, 7 July [online] https://developers.google.com/search/blog/2020/07/rich-results-test-out-of-beta (accessed 17 December 2020).

Schema.org (SDO) (2020) *Releases*, 4 December 2021 [online] https://schema.org/docs/releases.html (accessed 11 January 2021).

Schwardmann, U. (2020) 'Digital objects – FAIR Digital Objects: which services are required?', *Data Science Journal*, Vol. 19, No. 1, p.15 [online] http://doi.org/10.5334/dsj-2020-015.

Sefton, P., Carragáin, E.Ó., Soiland-Reyes, S., Corcho, O., Garijo, D., Palma, R. et al. (2020) *RO-Crate Metadata Specification 1.1 (Version 1.1.0)*, Zenodo, 30 October [online] http://doi.org/10.5281/zenodo.4031327.

Stocker, M. (2017) *Semantic Representation of Marine Monitoring Data*, April [online] https://www.researchgate.net/profile/Markus_Stocker/publication/317063084_Semantic_Representation_of_Marine_Monitoring_Data/links/5948daf6aca272f02e0c045b/Semantic-Representation-of-Marine-Monitoring-Data.pdf (accessed 18 December 2020).

Trust, Z. (2018) *Facilitating the Adoption of the FAIR Digital Object Framework in Material Science*, October [online] https://www.nist.gov/programs-projects/facilitating-adoption-fair-digital-object-framework-material-science (accessed 10 February 2021).

W3C (2015) *Syntax and Processing Rules for Embedding RDF Through Attributes*, 17 March [online] https://www.w3.org/TR/rdfa-core/ (accessed 10 December 2020).

Welcome to Schema.org (2020a) 4 December [online] https://schema.org/ (accessed 17 December 2020).

Welcome to Schema.org (2020b) *Crosswalks* [online] https://rd-alliance.github.io/Research-Metadata-Schemas-WG/ (accessed 18 December 2020).

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) 'The FAIR guiding principles for scientific data management and stewardship', *Sci. Data*, Vol. 3, p.160018 [online] https://doi.org/10.1038/sdata.2016.18.

World Meteorological Organization (WMO) (1959) *International Meteorological Vocabulary*, WMO-No. 182 ed. [online] https://library.wmo.int/doc_num.php?explnum_id=4712.