Project Details

Name of project:

2020 Science-On-Schema.Org Validator (2020-SOSOV)

Project lead and contact details:

Dave Vieglais, Biodiversity Institute, University of Kansas, vieglais@ku.edu

Project partners and contact details:

Doug Fils, Consortium for Ocean Leadership, dfils@oceanleadership.org
Adam Shepard, BCO-DMO, Woods Hole Oceanographic Institution, ashepherd@whoi.edu
Matt Jones, NCEAS, University of California, Santa Barbara, jones@nceas.ucsb.edu

Proposed start and end date:

2020-02-01 through 2020-10-31

Budget Requested:

$7,000

Budget Summary:

Category	Description	Amount
Facilities	AWS service hosting for evaluation and production instances	$2,600
Travel	PI attendance at ESIP 2020 Summer meeting (Burlington, VT) to present research progress and output.	$1,900
	PI attendance at Biodiversity Summit 2020 (Alexandria, VA) to present and promote ESIP science-on-schema.org recommendations.	$1,300
	PI attendance at ESIP 2021 Winter meeting (Bethesda, MD) to present research progress and output.	$1,200
	Total	$7,000

Project Outline

Project description:

2020-SOSOV will provide a publicly accessible, programmatic implementation of the ESIP Science-on-schema.org guidelines[1] to support compliance testing for schema.org[2] Dataset publishers. It will also support a reference implementation of Gleaner[3] to demonstrate harvest and indexing of schema.org Datasets.

Schema.org is attractive to Internet content publishers, in part because it promises a light-weight mechanism for advertising resources such as Datasets in a way that is easily machine harvested and evaluated. Schema.org is widely adopted for describing Internet resources in general, and is becoming more popular as a mechanism for advertising availability of scientific datasets. Content indexers such as Google[4], [5] and DataONE[6] are leveraging this to further improve findability of relevant data resources.

However, the flexibility of schema.org implementation can result in many different approaches to describe the similar resources. This in turn makes access to those resources more challenging for consumers since it is necessary to first develop an understanding of how a particular resource has been described. In extreme cases, different providers may use the same terms or even classes for different, incompatible purposes, diminishing the value of the schema.org approach. Unfortunately, just because content is described by schema.org markup does not necessarily mean that it can be readily accessed and reused.

To address this, the ESIP Science-On-Schema.Org cluster is developing guidelines and best practices for implementers wishing to publish and retrieve datasets using schema.org. These guidelines include technical specifications that provide the information necessary for developers to implement their schema.org markup in a consistent manner across the community. This in turn simplifies access to, and reuse of those resources since a consumer can trust an implementation follows the guidelines and so the potential permutations that need to be supported are significantly reduced. This is especially important for aggregation of content from multiple sources, since customization for each provider can be very expensive to implement and maintain.

Experience with harvesting and indexing schema.org content with the DataONE infrastructure has shown that even with careful communication between developers, errors and inconsistencies in implementation can impede successful deployments. However, there is currently no available test suite that evaluates a schema.org endpoint and provides feedback on compliance with the ESIP Science-on-schema.org guidelines. Google provides a service for evaluating structured data[7], though the guidance is very general in nature. Systems like Gleaner[8] test input on ingest, though are not designed to provide iterative testing feedback to content providers. In order to simplify and streamline the process of schema.org implementation, DataONE has developed a testing tool that evaluates each stage of schema.org Dataset publication[9], [10]. This web based tool evaluates the sitemap, landing page construct, schema.org markup, and resources referenced therein and provide a report on inconsistencies discovered, though it is currently directed specifically to the requirements of DataONE.

This project will refactor the DataONE validation tool to utilize and implement the ESIP Science-on-schema.org guidelines. The Science-On-Schema.Org Validator (SOSOV) will accept as input a data repository URL, a sitemap XML URL, individual landing page URLs, or schema.org JSON-LD markup to provide different levels of evaluation. Output will be readily re-usable JSON format that can be inspected manually by developers, utilized by automated test suites, or rendered to HTML for easy human reading. Gleaner will also operate in parallel with the validation suite to provide indexed views and discovery capabilities so that contributors can evaluate how the final result appears, especially in comparison with other indexed resources.

SOSOV and Gleaner will be available as source as well as Docker[11] images for self hosted deployment if desired. Stable, production deployments of SOSOV and Gleaner will be deployed on Amazon EC2 to facilitate integration with continuous build and test environments such as TravisCI[12].

Project objectives, significance, and impact:

The principal objective of this project is to provide a publicly accessible service where data content providers can evaluate and receive technical feedback on how their schema.org content performs with respect to the ESIP Science-on-schema.org guidelines. Other objectives include providing feedback on guideline documentation based on test suite implementation and commonly observed issues, and supporting Gleaner, a reference implementation of a schema.org Dataset harvesting and indexing environment.

Provision of this service is significant because it provides an independent, readily available mechanism for testing both schema.org Dataset publishers and the ESIP Science-on-schema.org guidelines. Such a resource is currently not available.

The impact will be reliable, more easily implemented and tested schema.org Dataset services that perform consistently and according to community developed guidelines.

Description of key project steps and timeline:

February	Establish project workspace Identify and document product requirements Review existing implementations and identify changes to meet requirements Identify the Minimum Viable Product (MVP) characteristics

March - May	Establish development environment with test service deployment Implement MVP Establish production deployment Provide access to Docker images for self hosting
June - October	Gather feedback, and iterate on service Present outcomes and gather feedback at ESIP summer meeting Promote and gather feedback from Biodiversity Summit meeting
October +	Present outcomes at Winter ESIP meeting.

Description of additional funding currently supporting this work:

There are no additional funds providing direct support for this activity. All project participants receive salary from their respective institutions. Potential for additional grants to directly or indirectly support this project are being explored.

Outreach

What groups/audiences will be engaged in the project?

The principal audience of the project will be any person or group interested in following the ESIP Science-on-schema.org guidelines. This includes any data repository that does currently or plans to publish schema.org Dataset feeds. Dataset users will also benefit from the project as they could quickly determine if a particular repository conforms with the ESIP Science-on-schema.org guidelines, and so be confident their tools will be able to reliably access content from the repository.

How will you judge the project's impact?

The impact of the project will be determined directly from usage metrics gathered from the service, and indirectly through feedback provided by users of the service and during iterations of service releases. Another measure of impact will be the number of schema.org implementers that conform with the ESIP Science-on-schema.org guidelines, however this measure lacks a control and thus can be considered qualitative at best.

How will you share the knowledge generated by the project?

Project outcomes will be shared through the project website, the project and related GitHub repositories, through outreach in various relevant fora including social media and meetings, and through formal publication.

Description of who (agencies/individuals) should be aware of this project, i.e. potential outreach targets:

Any data repository interested in sharing data using the schema.org web publishing pattern may benefit from this project and so should be aware of it. Similarly, data users (consumers) may also benefit since they could pre-test a potential data source to determine compliance with the ESIP guidelines. Other communities such as represented by the Research Data Alliance (RDA) and Biodiversity Information Standards (TDWG) that are involved in the development of standards and guidelines for sharing earth science (and beyond) data may also benefit from awareness of the project both for the formal implementation of guidelines developed by the ESIP cluster, but also to augment those guidelines as necessary to support any additional requirements of their communities.

Project Partners (as applicable)

Description of project partners (agencies/individuals) and their involvement:

Dave Vieglais is a Senior Scientist at the Biodiversity Institute of the University of Kansas and the Director of Development and Operations for DataONE. Vieglais was involved in the design and

implementation of the prototype schema.org Dataset validation service being utilized by DataONE for potential member node evaluation and will be primarily responsible for all project activities. Doug Fils (Consortium for Ocean Leadership) and Adam Shepard (Technical Director, BCO-DMO, Woods Hole Oceanographic Institution) co-chair the ESIP Science-on-schema.org cluster and along with Matt Jones (Director of Informatics R&D for NCEAS, University of California, Santa Barbara and Director of DataONE) are actively engaged in guideline development as well as various schema.org provider and consumer implementations. Fils will continue to operate the Gleaner service in parallel with the validation service as part of this project. Jones, Shepard, and Fils will have principally advisory roles in the project, with further contributions as time and resources permit.

How will this project engage members of the ESIP community:

The outcome from this project is a publicly accessible, interactive service that provides feedback on the technical qualities of schema.org Dataset feeds for any repository. As such, any ESIP members (and beyond) are welcome to utilize the service and provide feedback on its utility. Providing a functional implementation of guidelines can help ensure a more interactive evaluation of those guidelines and the consequence of changes during development iterations. Project outcomes will be presented at the upcoming ESIP summer and winter meetings, and through regular updates to the ESIP Science-on-schema.org cluster.

References

[1] "Provides guidance for publishing schema.org as JSON-LD for the sciences: ESIPFed/science-on-schema.org," 15-Nov-2018. [Online]. Available: https://github.com/ESIPFed/science-on-schema.org. [Accessed: 11-Dec-2018].

[2] "Full Hierarchy - schema.org." [Online]. Available: https://schema.org/docs/full.html. [Accessed: 24-Oct-2018].

[3] *Harvesting code for P418*. EarthCube Architecture - Project 418, 2018.

[4] N. Noy and D. Brickey, "Facilitating the discovery of public datasets," *Google AI Blog*, 24-Jan-2017. [Online]. Available: http://ai.googleblog.com/2017/01/facilitating-discovery-of-public.html. [Accessed: 24-Oct-2018].

[5] N. Noy, "Making it easier to discover datasets," *Google*, 05-Sep-2018. [Online]. Available: https://www.blog.google/products/search/making-it-easier-discover-datasets/. [Accessed: 24-Oct-2018].

[6] "DataONE Data Catalog." [Online]. Available: https://search.dataone.org/data. [Accessed: 10-Dec-2018].

[7] "Structured Data Testing Tool." [Online]. Available: https://search.google.com/structured-data/testing-tool/u/0/. [Accessed: 10-Jan-2020].

[8] *Harvesting code for P418*. EarthCube Architecture - Project 418, 2018.

[9] "DataONE schema.org scanner." [Online]. Available: https://so.test.dataone.org/schema_org/. [Accessed: 10-Jan-2020].

[10] D. A. Vieglais, J. G. Evans, and R. Dahl, "IN22B-04 - An Open Source Tool for Test and Evaluation of Schema.org Dataset Publishing." [Online]. Available: https://agu2019fallmeeting-agu.ipostersessions.com/default.aspx?s=57-7F-8E-5E-BC-E7-31-37-A3-52-C9-D5-BB-7E-48-A3. [Accessed: 10-Jan-2020].

[11] "Empowering App Development for Developers | Docker," *Docker*. [Online]. Available: https://www.docker.com/. [Accessed: 10-Jan-2020].

[12] "Travis CI - Test and Deploy with Confidence." [Online]. Available: https://travis-ci.com/. [Accessed: 10-Jan-2020].