class: center, middle, inverse

Report on EarthCube CDF Registry Working Group

Work to Date

.footnote[created with remark]

Group Members

People	Institute	Facility
Time Ahern	IRIS	IRIS
Bob Arko	LDEO / Columbia	R2R
Doug Fils	Ocean Leadership	Open Core Data
Danie Kinkade	WHOI	BCO-DMO
Lynne Schreiber	EarthCube ESSO	
Adam Shepherd	WHOI	BCO-DMO
Shelley Stall	AGU	COPDESS
Mike Stults	IRIS	IRIS

Michael Witt	Purdue	re3data

Acknowledgement: This work is support by the NSF EarthCube Program and the EarthCube Science Support Office (ESSO) which provides infrastructure and administrative support for the group. We are a working group.. NOT a funded project.

Group work is found at https://github.com/fils/CDFRegistryWG and https://github.com/fils/contextBuilder

---

Objectives and Benefits

Objectives

1. Formalize a set of repository parameters-of-interest to CDF members.
2. Review the alignment of those parameters with re3data and COPDESS.
3. Develop strategies for CDF members to express/expose this information.
4. Develop a means to encode this schema in a machine readable format.
5. Demonstrate the use of schema.org for publishing and accessing this metadata.
6. Leverage re3data as a reference implementation for collecting and exposing this metadata.

Benefits

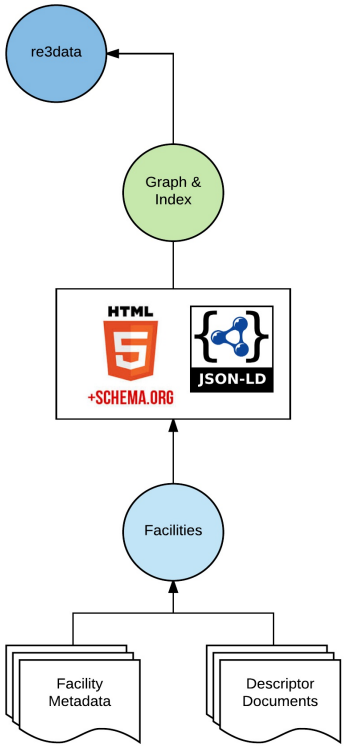
- Repositories have control over their metadata and can update at any time.
- These new standard guidelines for publishing repository metadata can be adopted across all ESS repositories and other scientific

domains to support repository discovery and access.
- CDF will recommend these standards to their membership and work towards adoption by all NSF-funded ESS repositories as complying with the standard.

Objectives and Benefits

Supporting data pipeline

- Encode in JSON-LD (schema.org and re3data vocabularies)
- Encode basic facility / repository metadata and links to service descriptions
- Harvest via simple whitelist crawl
- Generate graph and index and explore use of these products



Deliverables in support of pipeline

1. A draft ontology based on the re3data XML schema (https://github.com/fils/CDFRegistryWG/tree/master/Vocabulary) (A basis to encode extension review to re3 voc)
2. Table of possible extensions to the base re3data XML schema to support CDF facility description
3. A reference approach to implementing facility description using scehma.org type Organization and the developed ontology. (Reference example: https://github.com/fils/CDFRegistryWG/blob/master/opencore.json)

4. A crawler software package (repurposed) that extracts JSON-LD to RDF triples and generates a full text index from these crawls (https://github.com/fils/contextBuilder) See also https://github.com/fils/CDFRegistryWG/blob/master/onHarvesting.md (which is old and out of date)
5. A simple client to explore interface options on the graph and index. Client provides feedback on usability aspects of these products to aid assessment. (http://eccdf.cloudapp.net/)

Reference use case

---

re3data "harvest" mode

Facility exposes metadata about itself using JSON-LD embedded in its landing web page

Metadata to be harvested by re3data: the green elements above map directly to the re3data schema, orange elements will be extended using schema.org

Mock-up of extended facility information represented in the re3data web interface after it has been harvested

---

Reference use case

Publishing by providers: Self hosting metadata

Thank you all very much!

Institute	People	Status and Link

R2R	Bob Arko	Published
Open Core Data	Doug Fils	Published
BCO-DMO	Adam Shepherd / Danie Kinkade	Published
IRIS	Mike Stults / Tim Ahern	Published
UNAVCO	Jim Riley / Chuck Meertens	Published
Open Topology	Vishu Nandigam / Christopher Crosby	Published

- Check the URLs for contact point if you want to discuss this
- Publishing patterns are evolving (would love some shared experiences)
- Additionally IEDA and WHOI have expressed interest in publishing a well

---

Reference use case

Publishing by providers: Self hosting metadata

Secondary benefit: organic search benefits

Authority, Relevance and Trust

The approach has benefit beyond EarthCube's goals. It is an approach leveraging standards based approaches to organization description. It's arguably both a metadata publishing and outreach activity.

It is a path to address "semantic search optimization" for organizations.

This secondary benefit has not been assessed.

Effort required by Providers

At the current time the effort is relatively small since it scope only facility data normally found in a top page of a domain. Document contains links, if any, to existing service description documents.

A single page with a single JSON-LD document that can be hand crafted from templates (not scale-able).

.small[

```
<html lang="en">
<head>
    ...

    <script type="application/ld+json">
        {
            "@context": "http://schema.org/",
            "@type": "Organization",
            "name": "Open Core Data",
            "contactPoint": {
                "@type": "ContactPoint",
                ...
```

]

We are still refining the use and encoding of schema.org types and re3data terms (plus extensions)

Thoughts on scaling by providers

There is talk about how to scale

There are many good JSON-LD libraries and tools: https://json-ld.org/ Google and others also have structured data testing tools. reference

Leverage existing data pipelines the data facilities are doing now. Shared experiences via EarthCube, ESIP, others...

Some working group members (Open Core Data, BCO-DMO) are currently generating schema.org (JSON-LD) for our dataset landing pages. Connecting to these via type DataCatalog is being talked about.

Leverage our current approaches to allow a simplified and portable pattern to be shared with other providers. This is obviously an area to engage the larger audience of people doing this similar pattern.

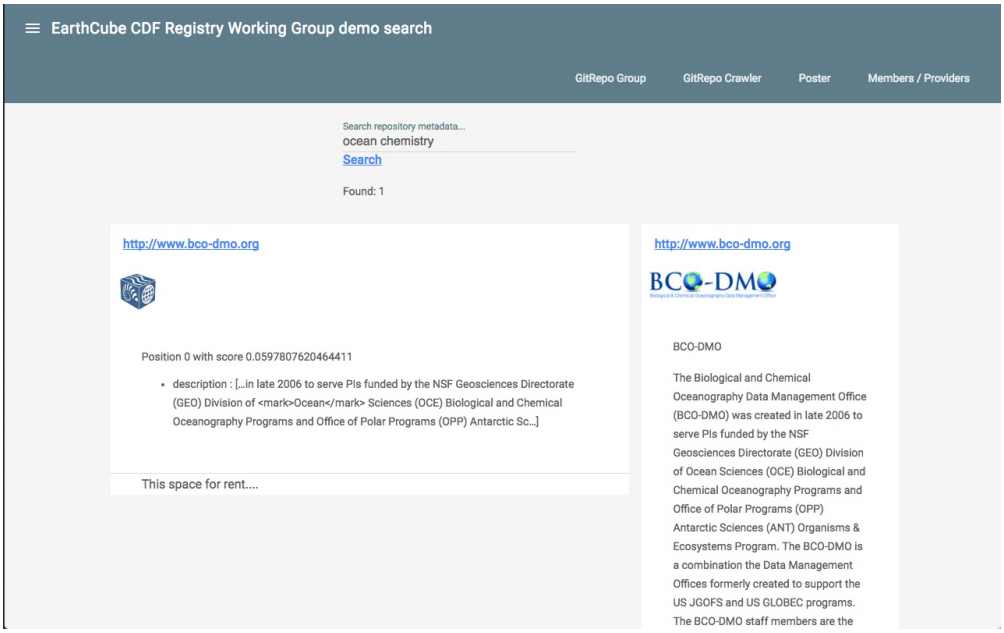
Harvesting

1. Harvest code (simple whitelist) contextBuilder:
2. Results in triples from JSON-LD and simple text index

The code is re-purposed from a previous abandoned experiment. So for a package with "context" in the name.. there is very little context.

Search and visualize

1. The resulting graph is TINY.. so it's mostly done by simple inspection with tools like Gephi, Cayley and Cytoscape
2. Simple search UI to exercise the index and graph. http://eccdf.cloudapp.net/?q=ocean+chemistry



Thoughts on work to date

Group will provide a report back to CDF at Summer ESIP (2017)

To support integration with schema.org the ontology development does have to take on certain approaches. These approaches facilitate an external ontology connecting to schema.org types. An “external vocabulary” is one of three approaches to working with scheme.org. All three approaches are valid and may fit various goals and situations. Ontologies can be used in JSON-LD without being a “external vocabulary” in schema.org space.

Use of schema.org + extensions is a viable approach. However, it does take a level of governance to ensure term coverage and mapping is taking place and well documented for both providers and consumers.

JSON-LD harvesting works but there are approaches to hypermedia navigation of the interconnected JSON-LD graph fragments that could make it simpler. Rather than employing whitelists or web crawling approaches. These include potential use of Hyda vocabulary or JSON-LD fragments.

The resulting graph is heavy with blank nodes under typical authoring and publishing approaches to JSON-LD. These may incur an impedance to using the resulting graph is a fully Linked Open Data approach. However, the situation can be addressed with authoring policy and guidelines and may not be a huge LOD publishing pattern issue in practice.

The graph and index are a product that can be used by any 3rd party. This might include the in group use case

represented by re3data or another group like Cinergi or DataOne. However, text indexes are very package specific and this project is not using the popular Solr index. Rather another package (bleve) was used that was in line with coding skills present in the group. It's likely it would be better to simply cache the JSON-LD documents (easy) and allow 3rd parties to build their own indexes from this cache.

The approach benefits both structured domain specific indexing as well as larger organic search systems (ala Google, Bing, Yandex, etc)

Future

- Complete pilot effort to test guidelines.
- Coordinate with schema.org as a potential external vocabulary for repository metadata for on-going management and governance.
- Provide guidelines to CDF members and other ESS NSF repositories to embed metadata information as JSON-LD (schema.org).
- Monitor and encourage adoption by CDF members and ESS NSF repositories

---

Next steps

- Exploring extending the connection down to DataCatalog (then to DataSet from there)
- Exploring use of schema.org proposed types: MeasurementTechnique and VariablesMeasured (leverage CSV for the Web patterns)
- JSON-LD walking leveraging JSON-LD Framing https://github.com/ESIPFed/snapHacks/tree/master/sh01-jsonldCrawl/simpleCrawler
- Making products and patterns more usable by a wider range of value add services
- more...

Thanks

Contact us

https://github.com/fils/CDFRegistryWG/blob/master/members.md

Public reports of group work:

- EarthCube All Hands June 2017 Poster
- ESIP Semantic Committee June 2017 Presentation
- DataONE Members Meeting July 2017 Poster
- EarthCube CDF July 2017 Report out at ESIP Summer 2017