Schema.org, Data repositories, and Federated Search

Ruth Duerr, Douglas Fils, Stephen Richard, Adam Shepherd, Ilya Zaslavsky
So, What’s the Problem?

- Polar data is highly distributed and extremely diverse
- This inhibits universal data discovery and access
Technologies enabling a possible solution?

- Large scale web crawling & harvesting
- Simple semantic technologies
What is Schema.org?

- Global collaborative community promoting structured data on the web
- Extensible vocabularies describing entities, actions, relations
- Widely used on millions of sites
- Founded by Google, Microsoft, Yahoo, and Yandex
- Community forum hosted by W3C
  - https://schema.org/
  - https://www.w3.org/community/schemaorg/
  - https://github.com/schemaorg/
**Why use it?**

- Simple path to adoption
- Lightweight, extensible vocabulary
- Generate from richer standards (ISO19139, EML, etc)
- We can teach Google, etc. about data
- Lingua franca for sharing of structured knowledge on the web
Process Overview

Metadata publisher

1. Map metadata to Schema.org
2. Embed JSON-LD in landing pages
3. Create & register sitemap

For details/demo attend
Wed. PM semantics session

Aggregator

1. Crawl/Harvest pages
2. Enrich & index
3. Expose via discovery portal

For details/demos attend
Thurs. AM semantics session
But will it solve all our data discovery and access issues?
A few Activities in this Space

1. bioschemas.org
2. NSF’s EarthCube’s GeoCODES projects
3. Earth Science Information Partners Federation Schema.org cluster
4. RDA Research Metadata Schemas Working Group
5. Google dataset search
A few Activities in this Space

1. bioschemas.org
2. NSF’s EarthCube’s GeoCODES projects
3. ESIP Federation Schema.org cluster
4. RDA Research Metadata Schemas Working Group
5. Google dataset search
What is Bioschemas?

Bioschemas aims to improve the Findability of data in the life sciences. It does this by encouraging people in the life sciences to use Schema.org markup in their websites so that they are indexable by search engines and other services. Bioschemas encourages the consistent use of markup to ease the consumption of the contained markup across many sites. This structured information then makes it easier to discover, collate, and analyse distributed data.

Bioschemas is making two main contributions:

1. Proposing new types and properties to Schema.org to allow for the description of life science resources.
2. Profiles over the Schema.org types that identify the essential properties to use in describing a resource.

Bioschemas started as a community effort in November 2015. It operates as an open community initiative with representatives from a wide variety of institutions. You are welcome to join the community.
Schema.org in Life Sciences

Bioschemas
bioschemas.org

BioThings
biothings.io

Customization

Health care image, modified from https://www.kissclipart.com/games-for-health-clipart-health-care-health-system-7qkz6d/
Bioschemas image: https://bioschemas.org/
BioThings image: https://biothings.io/
Customization

• Defining compulsory levels (all is optional in schema.org)

• Defining cardinality (all is 0 or more in schema.org)

• Recommending specialized vocabularies to be used as property objects (a lot of free text in schema.org)
Bioschema Types and Profiles

Types currently being integrated:

- BioChemEntity
- BioSample
- ChemicalSubstance
- Gene
- MolecularEntity
- Protein
- Taxon

Other types are still in draft mode

Profiles provide additional constraints beyond schema.org recommendations including:

- Information properties agreed by the community
  - minimum (M),
  - recommended (R),
  - optional (O),
- Property cardinality, i.e. whether it is expected to occur once or many times
- Associated controlled vocabulary terms drawn from existing ontologies (e.g., from Darwin Core)
  - Example: Taxon has properties taxonRank, childTaxon, parentTaxon
A few Activities in this Space

1. bioschemas.org

2. NSF’s EarthCube’s GeoCODES projects

3. ESIP Federation Schema.org cluster

4. RDA Research Metadata Schemas Working Group

5. Google dataset search
An Emerging Platform for Resource Registration, Discovery, and Access Utilizing Schema.org and Geoscience Vocabulary Extensions
A few Activities in this Space

1. bioschemas.org
2. NSF’s EarthCube’s GeoCODES projects
3. ESIP Federation Schema.org cluster
4. RDA Research Metadata Schemas Working Group
5. Google dataset search
ESIP Vision

Leaders in promoting the **collection, stewardship and (re)use**
Of Earth science data, information and knowledge
that is responsive to societal needs.
Dataset Discovery

Google dataset discovery
‘Facilitating Dataset Discovery using new developments within schema.org’ @ Summer ESIP 2013, Peter Fox & Doug Newman
7/2013

ESDSWG Search Relevance WG
Continued to discuss and move forward
Chris Lynnes, Lewis McGibbney
2015

P418
Separate NSF-related work for data facilities to implement dataset schema.
Doug Fils, Adam Shepherd, Eric Lingerfelt
2017

Google Dataset Search is finally live
9/2018

2014
NASA RDFa Mark-up carried over from ECHO to CMR

2016
ESIP Semantic Tech
Provides guidance for publishing schema.org as JSON-LD for the sciences
Beth Huffer, Doug Fils, Lewis,

1/2018
Google Dataset Search
Finally Google is utilizing Dataset schema for search; Google showed up at ESIP Winter Meeting 2018

Describing a Dataset

- Describing a Dataset
  - Common Properties
    - Identifier
    - Variables
    - Catalog
    - Metadata
    - Distributions
      - Accessing Data through a Service Endpoint
  - Temporal Coverage
  - Spatial Coverage
  - Roles of People
  - Publisher / Provider
  - Funding
  - Advanced Publishing Techniques
    - Attaching Physical Samples to a Dataset
A few Activities in this Space

1. bioschemas.org
2. NSF’s EarthCube’s GeoCODES projects
3. ESIP Federation Schema.org cluster
4. RDA Research Metadata Schemas Working Group
5. Google dataset search
WG Objective

- To identify and bridge gaps in existing schemas commonly used for research data, by bringing together communities who are working with such vocabularies to document research data and related resources;

- To provide guidelines for those communities whose needs are not addressed by existing metadata schema such as schema.org, and provide guidelines on proposing extensions.
A few Activities in this Space

1. bioschemas.org
2. NSF’s EarthCube’s GeoCODES projects
3. ESIP Federation Schema.org cluster
4. RDA Research Metadata Schemas Working Group
5. Google dataset search
Google’s “aims to support a strong open data system by encouraging”

1. Widespread adoption of open metadata formats to describe published data.

2. Further development of open metadata formats to describe more types of data and in more detail.

3. The culture of citing data the way we cite research publications, giving those who create and publish the data the credit that they deserve.

4. The development of tools that leverage this metadata to enable more discovery or better use of data. »
Google’s general data search capability

Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.
Google’s Search Results

Things to note:

1. 3 harvested records?
Things to note:

1. 3 harvested records?
2. Wow people are actually starting to cite data!
**Google’s Search Results**

**Things to note:**

1. 3 harvested records?
2. Wow people are actually starting to cite data!
3. Now who is the principle investigator?
1. Intended to cover all forms of data
Google’s Search Results Issues

1. Intended to cover all forms of data
2. Need to provide a uniform and predictable user experience
   1. They use general field names to display values from several fields (e.g. “provided by” to cover both publishers and creators)
   2. Do not use fields that are used differently by different sources
   3. Uses DOI’s and schema.org/sameAs to determine whether records are the same or not
Google’s Search Results Issues

1. Intended to cover all forms of data
2. Needs to provide a uniform and predictable user experience
   1. Use general field names to display values from several fields (e.g. provided by to cover publishers and creators)
   2. Do not use fields that are used differently by different sources
   3. Uses DOI’s and schema.org/sameAs to determine whether records are the same or not
3. Index new types of markup only when “enough” implementation has occurred to warrant it
Google’s Search Results Issues

1. Intended to cover all forms of data
2. Needs to provide a uniform and predictable user experience
   1. Use general field names to display values from several fields (e.g., provided by to cover publishers and creators)
   2. Do not use fields that are used differently by different sources
3. Uses DOI’s and schema.org/sameAs to determine whether records are the same or not
4. Not likely to support highly domain specific search criteria
Current Polar Harvesting Relationships Between Systems
1. Circular harvesting is all over the place!

2. Without a solid method of identifying the same record, eliminating duplicates might well be impossible!
Current Polar Harvesting Relationships Between Systems

1. Circular harvesting is all over the place!

2. Without a solid method of identifying the same record, eliminating duplicates might well be impossible!
So even Google has Issues Disambiguating Records!
But will it solve all our data discovery and access issues?
Only if Repositories and Aggregators agree on…

Semantics

Implementation

Repository A

Repository B

What we really need!