DATA DISCOVERY ENGINE

A Web-Based Toolset for Maximizing Data Discoverability and Promoting Reusable Data-Sharing Best Practices

Marco Cano  marcoc@scripps.edu
Xinghua (Jerry) Zhou  xzhou@scripps.edu
Use Cases

**Schema Playground**
(for developers)

- Visualize a Schema
- Create Your Own Schema by extending an existing class
- Schema Class Registry

**Discovery Best Practices**
(for data providers)

- Structured Metadata Discovery
- Dataset Metadata Registry

I have a new project and want to plan ahead so my data is findable, accessible, interoperable, and reusable (FAIR)!

I have a dataset and want to share it properly, so others can find it!
Schema.org provides a large number of classes in their schema we can use to structure and create a new schema for our specific purpose.

Too generic, does not fit all cases

Extend a class to create a new schema to fit your needs including your new class and its subclasses.
JSON-LD (Linked Data)

It is easy for humans to read and write. It is based on the already successful JSON format and it is the preferred format by Schema.org.

JSON Schema

Provides hints for user interfaces working with JSON data, and to make assertions about what a valid document must look like.

```
{  
  "@context": {  
    "@id": "http://schema.org/",  
    "schema": "http://schema.org/",  
    "bts": "http://discovery.biothings.io/bts/"  
  },  
  "@graph": [  
    {  
      "@id": "bts:BioMedicalDataset",  
      "@type": "rdfs:Class",  
      "rdfs:comment": "A body of structure information...",  
      "rdfs:label": "BioMedicalDataset",  
      "rdfs:subClassOf": {  
        "@id": "schema:Dataset"  
      },  
      "$validation": {  
        "$schema":  
        "http://json-schema.org/draft-07/schema#",  
        "type": "object",  
        "properties": {  
          "variableMeasured": {  
            "$comment": "...",  
            "description": "...",  
            "type": "enum"  
          },  
          "required": [  
            "variableMeasured"  
          ]  
        }  
      }  
    }  
  ]
}
```

Only needs to include new definitions and validation information.
Schema Playground - VISUALIZER

Class + Property Visualizer

Visualize and register your schema so others can use it for their research

Class + Schema Validation Visualizer

Eg. BioMedicalDataset

JSON-LD Schema

JSON-LD + JSON Schema Validation Schema

Eg. schema:Person, schema:Thing, etc..
A webtool and best practices guide to make sure your dataset metadata meets all the requirements by your needs and gets harvested by the data portals you are interested in.

**BioMedicalDataset**

Choose type of dataset schema you wish to follow

Follow guide and complete required and recommended fields by data portals you want to be discovered by.

A hosted page dedicated to your structured metadata

Tools to embed your structured metadata on your site.
One line of code to embed your structured metadata in your dataset page

```
<script src="https://discovery.biothings.io/api/dataset/83dc3401f86819de.js"></script>
```
Overview
Create Findable, Accessible, Interoperable, and Reusable Data

Create a Schema

Visualize Your Schema

Share Your Schema

+Schema Registry

Register Schema

Data Harvesting Portals
(Google dataset Search, CTSA Search, NIAID Data Portal, etc)

Get Found!

Create a Structured Metadata Based On Your Schema

Register Metadata

Thing

CreativeWork

Dataset

BioMedicalDataset
Acknowledgements:

Xinghua (Jerry) Zhou
Jiwen (Kevin) Xin
Sebastien Lelong
Chunlei Wu

Laura Hughes
Andrew Su

Sara Gonzales
Matthew B. Carson
Kristi L. Holmes

Check out our posters and learn more about our projects:

Structuring and crawling distributed biomedical metadata using schema.org standard
presenter: Xinghua Zhou
Poster Section P90

Data Discovery Engine Demo & Poster
presenter: Marco Cano
Poster Section P12

Learn more about our lab:
https://wulab.io

Try it: discovery.biothings.io

Thank You!

This project is developed as part of the CTSA National Center For Data To Health (CD2H), supported by grant U24TR002306 from National Center for Advancing Translational Sciences, National Institute of Health