Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions
Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions
Scraping data (perhaps using HTML)

HTML is a markup language to format documents for human readers.

It is now common to scrape data from the web, e.g., [Computerization of White Collar Jobs](https://www.upjohninstitute.org/) from UpJohn Institute, which scrapes online job postings.

Employer enters job info into database

User scrapes HTML to build local database

Database used to build HTML and post on web

Database analyzed by researcher
HTML is a markup language to format documents for human readers.

It is now common to scrape data from the web, e.g., Computerization of White Collar Jobs from UpJohn Institute, which scrapes online job postings.

Q: How make human readable document also machine readable?
A: Embed metadata into the HTML.
   For that we need standards for the metadata.
Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Health and medical types
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating
- Action
- JobPosting
- Dataset
Schema.org is already widely used

- See CACM article on [Schema.org: Evolution of Structured Data on the Web](#)
- ~45% of pages in the Google search index have schema.org markup
- Most popular mark up use
  - 6 billion product listings across 5m domains
  - 96m job postings (including duplicates)
- ...and many more from [comic books](#) to [recipes](#) to [mortgages](#), to [events](#) …
- Lesson: use schema.org standards if you want your data to be used
Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions
## JobPosting

A listing that describes a job opening in a certain organization.

### Properties from JobPosting

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>applicantLocationRequirements</td>
<td>AdministrativeArea</td>
<td>The location(s) applicants can apply from. This is usually used for telecommuting jobs where the applicant does not need to be in a physical office. Note: This should not be used for citizenship or work visa requirements.</td>
</tr>
<tr>
<td>applicationContact</td>
<td>ContactPoint</td>
<td>Contact details for further information relevant to this job posting.</td>
</tr>
<tr>
<td>baseSalary</td>
<td>MonetaryAmount or Number or PriceSpecification</td>
<td>The base salary of the job or of an employee in an EmployeeRole.</td>
</tr>
<tr>
<td>datoPosted</td>
<td>Date or DateTime</td>
<td>Publication date of an online listing.</td>
</tr>
<tr>
<td>educationRequirements</td>
<td>EducationalOccupationalCredential or Text</td>
<td>Educational background needed for the position or Occupation.</td>
</tr>
<tr>
<td>eligibilityToWorkRequirement</td>
<td>Text</td>
<td>The legal requirements such as citizenship, visa and other documentation required for an applicant to this job. A description of the employer, career opportunities and work environment for this position.</td>
</tr>
</tbody>
</table>
Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions
Dataset Search

Search for Datasets

Try coronavirus covid-19 or global temperatures.

Learn more about including your datasets in Dataset Search.
Proportion of Population Aged 65 or Older

Dataset updated Mar 31, 2016

Dataset provided by
(Point of Contact)

Description

Variable was created as part of a set of indicators that demonstrate links between the condition of natural areas and human concerns and that quantify dependencies on resources. More information about these resources, including the variables used in this study, may be found here:
Dataset

A body of structured information describing some topic(s) of interest.

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>distribution</td>
<td>DataDownload</td>
<td>A downloadable form of this dataset, at a specific location, in a specific format.</td>
</tr>
<tr>
<td>includedInDataCatalog</td>
<td>DataCatalog</td>
<td>A data catalog which contains this dataset. Supersedes catalog.</td>
</tr>
<tr>
<td>issn</td>
<td>Text</td>
<td>The International Standard Serial Number (ISSN) that identifies this serial publication. You can repeat this property to identify different</td>
</tr>
<tr>
<td></td>
<td></td>
<td>formats of, or the linking ISSN (ISSN-L) for, this serial publication.</td>
</tr>
<tr>
<td>measurementTechnique</td>
<td>Text or URL</td>
<td>A technique or technology used in a Dataset (or DataDownload), corresponding to the method used for measuring the corresponding variable(s) (described</td>
</tr>
<tr>
<td></td>
<td></td>
<td>using variableMeasured). This is oriented towards scientific and scholarly dataset publication but may have broader applicability; it is not intended as a full</td>
</tr>
<tr>
<td></td>
<td></td>
<td>representation of measurement, but rather as a high level summary for dataset discovery.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For example, if variableMeasured is: molecule concentration, measurementTechnique could be: &quot;mass spectrometry&quot; or &quot;nmr spectroscopy&quot; or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;colorimetry&quot; or &quot;immunofluorescence&quot;.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If the variableMeasured is &quot;depression rating&quot;, the measurementTechnique might be: &quot;self-report&quot; or &quot;clinician rating&quot;.</td>
</tr>
</tbody>
</table>
Table 3: Percentage of datasets with specific properties. Column 2 lists the source predicates for each property. Properties not listed in the table have values in fewer than 1% of the datasets.

<table>
<thead>
<tr>
<th>Property</th>
<th>Source predicates</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>so#description, purl#description</td>
<td>100.00%</td>
</tr>
<tr>
<td>title</td>
<td>so#name, purl#title</td>
<td>100.00%</td>
</tr>
<tr>
<td>provider</td>
<td>so#publisher, so#provider, purl#publisher</td>
<td>84.59%</td>
</tr>
<tr>
<td>keywords</td>
<td>so#keywords, dct#keyword, purl#keyword</td>
<td>80.08%</td>
</tr>
<tr>
<td>URL</td>
<td>so#url, dct#accessurl, dct#landigpage</td>
<td>68.30%</td>
</tr>
<tr>
<td>temporal coverage</td>
<td>so#temporalCoverage, so#temporal, purl#temporal</td>
<td>45.41%</td>
</tr>
<tr>
<td>data download</td>
<td>so#distribution, dct#distribution</td>
<td>44.34%</td>
</tr>
<tr>
<td>spatial coverage</td>
<td>so#spatialCoverage, so#spatial, purl#spatial</td>
<td>38.69%</td>
</tr>
<tr>
<td>date modified</td>
<td>so#dateModified, purl#modified</td>
<td>37.46%</td>
</tr>
<tr>
<td>license</td>
<td>so#license and so#license on so#distribution</td>
<td>34.80%</td>
</tr>
<tr>
<td>date published</td>
<td>so#datePublished, purl#published</td>
<td>30.83%</td>
</tr>
<tr>
<td>catalog</td>
<td>so#includedInCatalog</td>
<td>29.74%</td>
</tr>
<tr>
<td>variable</td>
<td>so#variableMeasured, dct#theme</td>
<td>20.90%</td>
</tr>
<tr>
<td>authors</td>
<td>so#author, so#creator</td>
<td>14.12%</td>
</tr>
<tr>
<td>same_as</td>
<td>so#sameAs, rdf#same_as</td>
<td>12.72%</td>
</tr>
<tr>
<td>date created</td>
<td>so#dateCreated</td>
<td>9.62%</td>
</tr>
<tr>
<td>alternate name</td>
<td>so#alternateName, rdf-schema#label</td>
<td>3.40%</td>
</tr>
<tr>
<td>is accessible for free</td>
<td>so#isAccessibleForFree</td>
<td>3.04%</td>
</tr>
</tbody>
</table>
Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions
We want to make it easier to find and analyze data

From search for dataset, download, clean, normalize, join ... to Just ask Google
Coming soon: In Google Search

- **Stats seeking queries** for $O(1000)$ variables from Demographics, Education, Employment, Health, Commerce, Disasters, Crime, and Housing
- **Line charts**
- **US data** from federal government; international data from World Bank, OECD
Important goal: Joining data across providers

Sites like FRED, Census, etc. allow joining data across their data sets by hosting the data and using common standards. But it is the common standards that really matter, since this allows joins across different distinct providers.

Start with dates, move on to geos.

E.g. I recently joined “county health rankings” (125 datasets) with “JHU covid confirmed cases by county” to build a predictive model at the county level.
Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions
Questions

1. What are best practices wrt to populating dataset metadata?
2. We are updating schema.org. What additional schemas do you want?
3. Who are important international partners?
4. Would “Job Trends” or “Recipe Trends” (modeled on Google Trends) be useful?