W3C Open Data Workshop

The schema.org experience

Dan Brickley <danbri@google.com>
Introducing schema.org

- Search engines collaborate to improve the Web
  - Bing, Yahoo!, Yandex and Google working together
  - Launched in June 2011
  - Large vocabulary for use in Web page markup
    - hundreds of types of thing & their properties
    - evolving via W3C-hosted open community
    - broad scope, practical focus
  - Based on W3C standards (RDFa, HTML5Microdata)
- Builds on ideas from Semantic Web, Microformats and Linked Data community
Simple example

New guidelines for metformin and diabetes mellitus

Dr. John Smith

Medical University

2012-03-24

We review clinical evidence related to the use of metformin for treatment of type-2 diabetes mellitus and provide new clinical guideline recommendations.

MeSH subject headings:

Metformin

Diabetes Mellitus, Type 2

Guidelines

Recommendation: we recommend monotherapy with metformin as an initial pharmacologic therapy to treat most patients with type 2 diabetes (Grade: strong recommendation).

high-quality evidence
Example uses at Google

- Enhanced search results - 'Rich Snippets'
- Custom Search Engines (for jobs, education, agriculture, datasets...)
- Working to improve description, discovery & integration of open datasets via schema.org
### Thing > CreativeWork > Article > NewsArticle

A news article

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>properties from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>additionalType</td>
<td>URL</td>
<td>An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. In RDFa syntax, it is better to use the native RDFa syntax – the 'typeof' attribute – for multiple types. Schema.org tools may have only weaker understanding of extra types, in particular those defined externally.</td>
</tr>
<tr>
<td>description</td>
<td>Text</td>
<td>A short description of the item.</td>
</tr>
<tr>
<td>image</td>
<td>URL</td>
<td>URL of an image of the item.</td>
</tr>
<tr>
<td>name</td>
<td>Text</td>
<td>The name of the item.</td>
</tr>
<tr>
<td>url</td>
<td>URL</td>
<td>URL of the item.</td>
</tr>
<tr>
<td>properties from</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CreativeWork</td>
<td></td>
<td></td>
</tr>
<tr>
<td>about</td>
<td>Thing</td>
<td>The subject matter of the content.</td>
</tr>
<tr>
<td>accountablePerson</td>
<td>Person</td>
<td>Specifies the Person that is legally accountable for the CreativeWork.</td>
</tr>
<tr>
<td>aggregateRating</td>
<td>AggregateRating</td>
<td>The overall rating, based on a collection of reviews or ratings, of the item.</td>
</tr>
<tr>
<td>alternativeHeadline</td>
<td>Text</td>
<td>A secondary title of the CreativeWork.</td>
</tr>
<tr>
<td>associatedMedia</td>
<td>MediaObject</td>
<td>The media objects that encode this creative work. This property is a synonym for encodings.</td>
</tr>
<tr>
<td>audience</td>
<td>Audience</td>
<td>The intended audience of the item, i.e. the group for whom the item was created.</td>
</tr>
<tr>
<td>audio</td>
<td>AudioObject</td>
<td>An embedded audio object.</td>
</tr>
<tr>
<td>author</td>
<td>Organization  or Person</td>
<td>The author of this content. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.</td>
</tr>
<tr>
<td>award</td>
<td>Text</td>
<td>An award won by this person or for this creative work.</td>
</tr>
<tr>
<td>awards</td>
<td>Text</td>
<td>Awards won by this person or for this creative work. (legacy spelling; see singular form, award)</td>
</tr>
<tr>
<td>comment</td>
<td>UserComments</td>
<td>Comments, typically from users, on this CreativeWork.</td>
</tr>
<tr>
<td>contentLocation</td>
<td>Place</td>
<td>The location of the content.</td>
</tr>
<tr>
<td>contentRating</td>
<td>Text</td>
<td>Official rating of a piece of content—for example, 'MPAA PG-13'.</td>
</tr>
<tr>
<td>contributor</td>
<td>Organization  or Persons</td>
<td>A secondary contributor to the CreativeWork.</td>
</tr>
</tbody>
</table>
This class contains derivatives of IPTC rNews properties. rNews is a data model of publishing metadata with serializations currently available for RDFa as well as HTML5 Microdata. More information about the IPTC and rNews can be found at [rnews.org](http://rnews.org).
### Proposals for Schema.org

**Accessibility**
- **Addiction candidate**
  - Extension for describing accessibility properties of a CreativeWork.

**Learning Resources**
- **Addiction candidate**
  - Draft based on LRM submission - please revisit.
  - The Learning Resources Metadata Initiative has drafted a set of terms to improve schema.org's use with educational materials.

**Datasets**
- **Addiction candidate**
  - A proposal extending schema.org for describing datasets and data catalogs.
  - Adds the DataSet, DataCatalog, and DataDownload types and supporting properties.
  - An example page with dataset metadata is available here.

**Type-based**
- **Addiction candidate**
  - A proposal extending schema.org for describing technical content.
  - Adds `technicalAudience` and `technicalLicenses`.

**AP reference context**
- **Addiction candidate**
  - A proposal extending schema.org for describing Application Programming Interface (API) reference context.
  - Derived from TechArtic; adds assembly, assemblyVersion, programmingModel, targetPlatform, etc.

**Code**
- **Addiction candidate**
  - A proposal extending schema.org for describing human-readable source code.
  - Adds `codeRepository`, `codemirror`, `programmingLanguage`, `targets`, `targetProduct`.

**Audience**
- **Addiction candidate**
  - A proposal extending schema.org for describing audiences, building around `http://schema.org/Audience` type, also used by (and coordinated with) LRM and MedicalVocabulary.
  - Add property `audienceType` to type `CreativeWork` (done) and `Product`. Also, `ParentsAudience`, ... see draft for details.
  - Partially included already; final tweaks.

**Paying and Actions**
- **Proposal**
  - A proposal for new classes Activity and Action, possibly also addressing some issues around UserInteraction.
  - Adds classes Activity, Action, AggregateActivity plus associated properties, and some recommendations on changes to related existing vocabulary.

**TV and Radio proposal**
- **Addiction candidate - please review!**
  - Proposes modest changes and additions to support TV and radio (from EBU and BBC).
  - Adds `Series`, `Episode`, `Season`, `Series/Season/Episode`, `RadioSeries`, `RadioSeason/RadioEpisode`, `RadioPodcast`, and modifications to the existing `Season/Episode` class.
  - Adds `Schedule/Episode` for programmes other than episodes.
  - Additional description of a service with `Service/PresentationEvent` for broadcast, streaming, etc.

**Sports**
- **Some initial discussions**
  - General sports are working towards a Sports Vocabulary proposal. See `http://schema.org`/Sports glossary.
  - See `proposal from Paul Ked (Graphic (PingPong))`. Microdata + Sports Stats post describing earlier ESPN/Google collaboration, and the Sports Ontology, used by BBC, Press Association, and others. See also [sports-schemas-changes mailing list](mailto:sports-schemas-changes@w3.org).

**Games and Sports**
- **Discussions on Proposals**
  - Proposal from Marvel (see example).
  - Adds `PeriodicSeries` (under `IntangibleThing`), `PeriodicalIssue` (under `CreativeWork`), Comic issue (under `PeriodicalIssue`), and `GraphicNovel` (under `Book`). Approx 8-9 properties for each of these classes.
  - Generally well received and qualified, but some detailed `discussion` re fine-grain detail of books, etc. Can those properties be shared with `Book/number/OPages`, `StabilityIssues`? Should `PeriodicSeries` be integral?

**Event schema**
- **Proposals and review**
  - Proposes some modifications to `Event` based on deployment experience.
  - Adds eventStatus, previousEndDate, previousEndDate, eventCategory, makes startDate and `dateRepeatable`.
  - Encourages use of `uri` from `Thing`.
  - Forty modest proposals based on implementor feedback. Some examples needed, especially around `startDateTime` (it's obvious how they work with multiple values! Still need work. Some pieces show up in a `Activities`/`Activities` proposal now.

**Historical Data**
- **Proposal**
  - The `historicaldata.org` site proposes some extensions to schema.org to address historical and genealogical (family history) scenarios.
  - Adds `HistoricalRecord`, `HistoricalEvent` (subclass of `Event`). Discusses issues, `subEvents`, `superEvent`, etc. Looks like the `PeriodicSeries` to point to `HistoricalEvent`.
  - There is an associated implementation, [Real History](http://www.realhistory.com), and [blog](http://blog.realhistory.com), but no implementation.

**Biological Database**
- **Proposal**
  - A schema extension for biological databases, `Genes`.
  - Adds a class `BiologicalDatabaseEntry`, a kind of `CreativeWork`, describing `entryID`, `entryType`, `genes`, `BiologicalDatabase` also subclass of `CreativeWork`, with no special properties. Also use `Breadcrumbs` from WebPage.

**Broadcrumbs**
- **Proposal**
  - A proposal resolving current problems handling broadcrumbs.
  - Adds `BreadCrumb` type with child property and suppose to use this class on breadcrumb property of a WebPart, an important schema, but needs attention.
  - There's an ongoing [discussion](https://www.w3.org/wiki/Schema.org/BreadcrumbsProposal) on using literal values.

**Discussion**
- **Proposal**
  - A proposal extending schema.org for describing online discussion forums and message boards.
  - Adds `attachmentURL`, `contributorType`, `discussionStatus`, `interestedUsers`, `postCount`, `postType`, `voteHelpful`.

**Innovation**
- **Proposal**
  - A proposal extending schema.org for describing innovation and related concepts. Discussion found here.
  - Covers concepts and terms such as `Named, Benefit, Usage, Disruption, Innovator, Innovation, Embodiment`, to name a few. See [Innovation](https://www.w3.org/wiki/Schema.org/Innovation) for complete list of terms.

**Proposals for Schema.org**

These proposals are working towards inclusion in changes.html, the file that serves as a log of changes, and for communicating requested changes to the schema.org administrators.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Status</th>
<th>Summary</th>
<th>Vocabulary</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessible</td>
<td>Addition candidate</td>
<td>Extension for describing accessibility properties of a CreativeWork.</td>
<td>Accessibility</td>
<td>Draft published; seek feedback.</td>
</tr>
<tr>
<td>Learning Resources</td>
<td>Addition candidate</td>
<td>Draft based on LRM submission - please revisit.</td>
<td>Financing</td>
<td>Comments, see version 1.0 and discussion list.</td>
</tr>
<tr>
<td>Datasets</td>
<td>Addition candidate</td>
<td>A proposal extending schema.org for describing datasets and data catalogs.</td>
<td>DataCatalog</td>
<td>An example page with dataset metadata is available here. For related efforts, see the WSC Government Linked Data Working Group.</td>
</tr>
<tr>
<td>Type-based</td>
<td>Addition candidate</td>
<td>A proposal extending schema.org for describing technical content.</td>
<td>Technical</td>
<td>Addition candidate</td>
</tr>
<tr>
<td>AP reference context</td>
<td>Addition candidate</td>
<td>A proposal extending schema.org for describing Application Programming Interface (API) reference context.</td>
<td>ProgrammingModel</td>
<td>Addition candidate</td>
</tr>
<tr>
<td>Audience</td>
<td>Addition candidate</td>
<td>A proposal extending schema.org for describing audiences, building around <code>http://schema.org/Audience</code> type, also used by (and coordinated with) LRM and Medical Vocabulary.</td>
<td>PeopleAudience</td>
<td>Partially included already; final tweaks.</td>
</tr>
<tr>
<td>Paying and Actions</td>
<td>Proposal</td>
<td>A proposal for new classes Activity and Action, possibly also addressing some issues around UserInteraction.</td>
<td>Activity</td>
<td>This addresses some known problems with the current UserInteraction design, and integrated activityобытиe functionality into schema.org.</td>
</tr>
<tr>
<td>TV and Radio proposal</td>
<td>Proposal</td>
<td>Proposes modest changes and additions to support TV and radio (from EBU and BBC).</td>
<td>ProgramEvent</td>
<td>See also background image and [11]; Draft initially from EBU/EBBC/TECtube discussions. Nearby vocab: <a href="http://schema.org/VideoObject">VideoObject</a>, <a href="http://schema.org/MusicObject">MusicObject</a>. Before re-scheduling. See also <a href="http://schema.org/UseBlog">UseBlog</a>, a kind of UserInteraction. Comics and TV/Radio also have a concern for describing fictional characters and narratives.</td>
</tr>
<tr>
<td>Sports</td>
<td>Some initial discussions</td>
<td>General sports are working towards a Sports Vocabulary proposal. See <code>http://schema.org</code>/Sports glossary.</td>
<td>SportEquipment</td>
<td>Several sources including a full draft from the ESPN/Google work. Needs summary here.</td>
</tr>
<tr>
<td>Games and Sports</td>
<td>Proposals</td>
<td>Discussion on Proposals</td>
<td>ComicIssue</td>
<td>March 2nd - post write-up of the schema.org Microdata vocabulary, next discussion is on defining common basic sports vocab for schema.org.</td>
</tr>
<tr>
<td>Event schema</td>
<td>Proposals and review</td>
<td>Proposes some modifications to <code>Event</code> based on deployment experience.</td>
<td>Event</td>
<td>Forty modest proposals based on implementor feedback. Some examples needed, especially around <code>startDateTime</code> (it's obvious how they work with multiple values! Still need work. Some pieces show up in <code>Activities</code>/<code>Activities</code> proposal now.</td>
</tr>
<tr>
<td>Historical Data</td>
<td>Proposal</td>
<td>The <code>historicaldata.org</code> site proposes some extensions to schema.org to address historical and genealogical (family history) scenarios.</td>
<td>HistoricalRecord</td>
<td>There is an associated implementation, <a href="http://www.realhistory.com">Real History</a>, and <a href="http://blog.realhistory.com">blog</a>, but no implementation.</td>
</tr>
<tr>
<td>Biological Database</td>
<td>Proposal</td>
<td>A schema extension for biological databases, <code>Genes</code>.</td>
<td>BiologicalDatabase</td>
<td>Others have also mentioned interest in adding some notion of species.</td>
</tr>
<tr>
<td>Broadcrumbs</td>
<td>Proposal</td>
<td>A proposal resolving current problems handling broadcrumbs.</td>
<td>BreadCrumb</td>
<td>There's an ongoing discussion on using literal values.</td>
</tr>
<tr>
<td>Discussion</td>
<td>Proposal</td>
<td>A proposal extending schema.org for describing online discussion forums and message boards.</td>
<td>AttachmentURL</td>
<td>Addition candidate</td>
</tr>
<tr>
<td>Innovation</td>
<td>Proposal</td>
<td>A proposal extending schema.org for describing innovation and related concepts.</td>
<td>Innovation</td>
<td>Addition candidate</td>
</tr>
<tr>
<td>Proposal for long</td>
<td>How far along</td>
<td>Summary of contributions</td>
<td>Overview of new terms</td>
<td>Comments on status, particularly regarding integration, overlaps and open issues.</td>
</tr>
</tbody>
</table>
Datasets

- Please read our position paper!
- We've added types for Dataset, DataDownload, DataCatalog
  - thanks to Jim Hendler & team, W3C DCAT, ...
- Allows any/all schema.org properties to be used for dataset discovery
  - but what about the information in the data? can we find ways to look inside datasets?
- Where do graphs stop and tables start?
  - fulfilling RDF's original promise as a metadata system
- Collaborations: W3C Community groups?