deepschema.org: An Ontology for Typing Entities in the Web of Data

Panayiotis Smeros, Amit Gupta, Michele Catasta and Karl Aberer
What is deepschema.org?

- Class Hierarchy, Taxonomy

- Describes all the possible environments than an entity exists in

- Comprises information from Wikidata and schema.org

- Main features:
  - generic, cross-domain
  - rich, evolving as fast as the Web
  - traversable
  - accurate
Related Work

**YAGO Taxonomy**
- ✓ rich, evolving as fast as the Web
- × traversable
- ✓ accurate

<wikicat_People_murdered_in_British_Columbia> rdfs:subClassOf <wordnet_person_100007846>.

**Dbpedia Ontology**
- × rich, evolving as fast as the Web
- ✓ traversable
- ✓ accurate

Manually constructed; Only 685 classes

**Wikidata Class Hierarchy**
- ✓ rich, evolving as fast as the Web
- ? traversable
- ✓ accurate

Crowdsourced schema; No tree structure

**schema.org**
- ? rich, evolving as fast as the Web
- ✓ traversable
- ✓ accurate

Manually constructed; Used by billions of web pages

deepschema.org: An Ontology for Typing Entities in the Web of Data
Wikidata Class Hierarchy
Wikidata Class Hierarchy

Extraction Phase

RDFS Entailment Rules

A rdf:type B ⇒ class(B)
A rdfs:subClassOf B ⇒ class(A) ∧ class(B) ∧ subclass(A, B)

Filtering Phase

Ontologies from domain-specific KBs

Classes with no English label

Keep deepschema.org generic

A rdfs:subClassOf ≡ P279 (subclass of)
rdfs:type ≡ P31 (instance of)
## Structure of the Class Hierarchy

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>123,033</td>
</tr>
<tr>
<td>Subclass Relations</td>
<td>126,688</td>
</tr>
<tr>
<td>Disconnected Subgraphs</td>
<td>4,263</td>
</tr>
<tr>
<td>Root Classes</td>
<td>14,084</td>
</tr>
<tr>
<td>Leaf Classes</td>
<td>102,434</td>
</tr>
<tr>
<td>Avg num of Subclasses per Class</td>
<td>1.03</td>
</tr>
<tr>
<td>Avg depth of Hierarchy</td>
<td>7.93</td>
</tr>
</tbody>
</table>

One subgraph contains **96%** of the total classes and **97%** of the total subclass relations.
The first class with more **direct** than **inherited** instances is **human**.
For a *non-English* class hierarchy we get *at least* ~50% loss of information.
Number of Classes per Wikidata External Contributor

- Freebase
- English Wikipedia
- German Wikipedia
- German National Library
- French Wikipedia
- Library of Florence
- ... dbpedia
- schema.org
- no provenance

50% of the classes has *no provenance* information
Integration with schema.org
Integration Heuristics

- **Exact Match**
  \((\text{language})^w \rightarrow (\text{Language})^s\)

- **Lemma Match**
  \((\text{languages})^w \rightarrow (\text{Language})^s\)

- **Head Match**
  \((\text{Kalapuyan languages})^w \rightarrow (\text{Language})^s\)

- **Vector Cosine Similarity**
  \((\text{warehouse})^w \rightarrow (\text{Store})^s\)

- **Head Similarity**
  \((\text{Survey motor boat})^w \rightarrow (\text{Vessel})^s\)

- **Subclass/Instance Similarity**
### Resulting pairs

<table>
<thead>
<tr>
<th>Cosine Similarity Threshold</th>
<th># of pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>15,112</td>
</tr>
<tr>
<td>0.6</td>
<td>8,494</td>
</tr>
<tr>
<td>0.7</td>
<td>7,329</td>
</tr>
<tr>
<td>0.8</td>
<td>6,120</td>
</tr>
<tr>
<td>0.9</td>
<td>5,586</td>
</tr>
</tbody>
</table>
Evaluation

(Accuracy, Traversability and Genericity)
Accuracy: Crowdsourcing Evaluation

- CrowdFlower Platform
- ~100 workers
- Majority Voting (2 out of 3)

<table>
<thead>
<tr>
<th>Class: Google driverless car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description: project by Google that involves developing technology for driverless cars</td>
</tr>
<tr>
<td>Link: <a href="http://www.wikidata.org/wiki/Q15330">http://www.wikidata.org/wiki/Q15330</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation: SubClassOf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class: Car</td>
</tr>
<tr>
<td>Description: A car is a wheeled, self-powered motor vehicle used for transportation.</td>
</tr>
<tr>
<td>Link: <a href="http://schema.org/Car">http://schema.org/Car</a></td>
</tr>
</tbody>
</table>

Is the relation valid?
- Yes
- No
Accuracy: Crowdsourcing Evaluation

**WikiData**: 92%

**schema.org**: 100% (assumed)

- **YAGO (Wikipedia Categories - WordNet)**: 95%
- **PARIS (YAGO - DBpedia)**: 90%

Comparable with integration techniques for similar data sources

Integration layer

![Graph showing accuracy percentages](chart)

- 73% at 0.5
- 85% at 0.6
- 88% at 0.7
- 91% at 0.8
- 91% at 0.9
Traversability

- **False class definition**
  - `partOf` relations interpreted into `subclassOf`

- **Topic not covered by schema.org**
  - Class `Child Abuse`
Genericity

Oxford English Dictionary

3000 most frequent English words

Nouns and Noun Phrases

81% coverage
Conclusions & Future Work

- Integrate more data sources
  - Facebook’s Open Graph

- Include customizable filters
  - Control the granularity of the ontology
  - Choose specific topics of interest

- Employ deepschema.org in use-cases
  - Discover the appropriate type (class) of an entity, given a context

deepschema.org

✔ generic, cross-domain
✔ rich, evolving as fast as the Web
✔ traversable
✔ accurate
Thanks for your attention!
Questions?