Microdata and schema.org

What is WHATWG?

- Web Hypertext Application Technology Working Group
  - Community interested in evolving the Web with focus on HTML and Web API development
  - Ian Hickson is a key person, now at Google
- Founded in 2004 by individuals from Apple, Mozilla and Opera after a W3C workshop
  - Concern about W3C's embrace of XHTML
- Current work on HTML5
- Developed Microdata spec

Basics

- Microdata is a simple semantic markup scheme that's an alternative to RDFa
- Developed by WHATWG and supported by major search companies (Google, MSFT, Yahoo)
- Like RDFa, it uses HTML tag attributes to host metadata
- Vocabularies are controlled and hosted at schema.org

http://whatwg.org/
**HTML5**

- Started by WHATWG as an alternative to XHTML, joined by W3C
  - A W3C candidate recommendation in 2012
  - WHATWG will evolve it as a “living standard”
- HTML5 ≈ HTML + CSS + js
- Native support for graphics, video, audio, speech, semantic markup, …
- Partial support in current browsers + extensions

**Microdata**

- The microdata effort has two parts: markup and a set of vocabularies
- The markup is similar to RDFa in that it provides a way to identify subjects, types, properties and objects
- The sanctioned vocabularies are found at schema.org and include a small number of very useful ones: people, movies, etc.

**HTML taxonomy and status**

**An example**

```html
<div>
  <h1>Avatar</h1>
  <span>Director: James Cameron (born 1954) </span>
  <span>Science fiction</span>
  <a href="avatar-trailer.html">Trailer</a>
</div>
```
An example: itemscope

- An `itemscope` attribute identifies a content subtree that is the subject about which we want to say something

```html
<div itemscope>
  <h1>Avatar</h1>
  <span>Director: James Cameron (born 1954)</span>
  <span>Science fiction</span>
  <a href="avatar-trailer.html">Trailer</a>
</div>
```

An example: itemtype

- An `itemscope` attribute identifies a content subtree that is the subject about which we want to say something
- The `itemtype` attribute specifies the subject’s type

```html
<div itemscope itemtype="http://schema.org/Movie">
  <h1>Avatar</h1>
  <span>Director: James Cameron (born 1954)</span>
  <span>Science fiction</span>
  <a href="avatar-trailer.html">Trailer</a>
</div>
```

An example: itemprop

- An `itemscope` attribute identifies a content subtree that is the subject about which we want to say something
- The `itemtype` attribute specifies the subject’s type
- An `itemprop` attribute gives a property of that type

```html
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
    Director: <span itemprop="name">James Cameron</span>
    (born <span itemprop="birthDate">1954</span>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="avatar-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

An example: embedded items

- An itemprop immediately followed by another itemprop makes the value an object

```html
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemtype="http://schema.org/Person">
    Director: <span itemprop="name">James Cameron</span> (born <span itemprop="birthDate">1954</span>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="avatar-trailer.html" itemprop="trailer">Trailer</a>
</div>
```
schema.org vocabulary

- Full type hierarchy in one file
- As of 4/23/13: 419 classes, 756 properties
- **Data types**: Boolean, Date, DateTime, Number (Float, Integer, Text (URL), Time
- **Objects**: Rooted at Thing with two ‘metaclasses’ (Class and Property) and eight subclasses

http://schema.org

http://schema.rdf.org

**Microdata as a KR language**

- More than RDF, less than RDFS
- Properties have an *expected* type (range)
  - Might be a string
  - A list of types, any of which are OK
- Properties attached to one or more types (domain)
- Classes can have multiple parents and inherit (properties) from all of them
- No axioms (e.g., disjointness, cardinality, etc.)
Mixing markup from other vocabularies

- Microdata is intended to work with one vocabulary – the one at schema.org
- Advantages
  - Simple, organized, well designed
  - Controlled by the schema.org people
- Disadvantages: too simple, controlled
  - Too simple, narrow, mono-lingual
  - Controlled by the schema.org people

Extending the schema.org ontology

- http://www.schema.org/docs/extension.html
- You can subclass existing classes
  - Person/Engineer
  - Person/Engineer/ElectricalEngineer
- Subclass existing properties
  - musicGroupMember/leadVocalist
  - musicGroupMember/leadGuitar1
  - musicGroupMember/leadGuitar2

Extension Problems

- Do agreed upon meaning
  - Through axioms supported by the language (e.g., equivalence, disjointness, etc.)
  - No place for documentation (annotations, labels, comments)
- Without a namespace mechanism, your Person/Engineer and mine can be confused and might mean different things

Conclusions

- Microdata is a good effort by the search companies to experiment with a simple semantic language
- It’s not a great standard
- RDFa has a more powerful encoding and works with the RDF stack
- There’s a bit of infighting in the WEB community
- RDFa Lite is maybe a good solution