INFO116

Adding Semantics to Data

Data and Semantics

- Textbook: Chapter 3
- What kind of data are you dealing with?
  - Unstructured
  - Semi structured
  - Structured (Ontology)
  - Mutimedia
- How to find/add the semantic metadata?

Unstructured data

- Grammatical text (most web sites - blogs - wikis)
- Application specific, user generated content

Grammatical text

- Newspapers, journals, magazines, books, web sites
- full sentences satisfying grammatical constraints (more or less)
- No technology that can “understand” text at a level of a human reader
  - named entity extraction, extraction of relationships, and “understanding” a restricted set of inputs (e.g. Siri)

OrgName	LocationName
Omnicom	New York
DDB Needham	New York
Kaplan Thaler Group	New York
BBDO South	Atlanta
Georgia-Pacific	Atlanta

Text extraction

- The fourth Wells account moving to another agency is the packaged paper-products division of Georgia-Pacific Corp., which arrived at Wells only last fall. Like Hertz and the History Channel, it is also leaving for an Omnicom-owned agency, the BBDO South unit of BBDO Worldwide. BBDO South in Atlanta, which handles corporate advertising for Georgia-Pacific, will assume additional duties for brands like Angel Soft toilet tissue and Sparkle paper towels, said Ken Haldin, a spokesman for Georgia-Pacific in Atlanta.

Companies that operate in Atlanta

OrgName
BBDO South
Georgia-Pacific

Application specific user generated content

- Twitter, Facebook
- terse, colloquial, context-poor, and containing words/phrases with nonstandard/creative, homophone-based spellings
- assimilation of content requires background knowledge
  - spatio-temporal-thematic context
- novel techniques and tools to summarize and visualize aggregated content, e.g. sentiment analysis

Twitter sentiment

http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Semi structured data

- XML, HTML
- tag makes explicit the category and the properties of the enclosed text using attribute-value pairs
- If you know the definition of the tags, you can extract some semantics
  - e.g. “important” words could be all words wrapped in <h1>, <h2>, etc.

Structured data

- Well defined formal syntax + Associated data model
  - Relational database tables
  - Sensor data stream
  - Web service call
- Easily parsed
- Must have access to data definition model

Techniques for adding semantics to data

- Microformats
- RDFa
  - OGP (Facebook)
- Microdata
- RDFa Lite
- JSON-LD

Microformats

- Reuse existing HTML/XHTML tags to incorporate metadata
- Embeds and encodes semantics within the attributes of markup tags
- Both human readable and machine processable
- Microformats is a community driven effort to extend semantics to well specified, precise domains

Microformats

- hCalendar - events
- hCard - people, organizations, contacts
- rel-license - licensed content
- rel-nofollow - links in untrusted 3rd party content
- rel-tag - tag posts and pages by subject
- XFN - social relationships and rel-me links among profiles for the same person
- XMDP - define a microformat vocabulary / profile
- XOXO - outlines

hCard example

Amit Sheth </ div >

Kno.e.sis Center 937–775–5217 http :// knoesis .org/

</ div >

can be semantically enhanced using the hCard microformat markup as

<div class=”card">

Amit Sheth

Kno.e.sis Center

937–775–5217

http://knoesis.org/

</ div >

RDFa

- Resource Description Framework - in - attributes
- Set of attribute-level extensions to XHTML
  - embedding rich metadata within Web documents
- Improves traceability and minimizes duplication of information

MICRODATA

- Proposed HTML5 specification used to embed semantics within Web documents
- Simpler than RDFa and Microformats
- Google, schema.org
  - a schema / simple ontology for web resources

schema.org with Microdata

```
<div>
 <h1>Avatar</h1>
 <span>Director: James Cameron (born August 16, 1954)</span>
 <span>Science fiction</span>
 <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

```
<div itemscope>
  <h1>Avatar</h1>
  <span>Director: James Cameron (born August 16, 1954) </span>
  <span>Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1>Avatar</h1>
  <span>Director: James Cameron (born August 16, 1954)</span>
  <span>Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html">Trailer</a>
</div>
```

```
<div itemscope itemtype ="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <span>Director: <span itemprop="director">James Cameron</span> (born August 16,
1954)</span>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

schema.org: Movie

Thing > CreativeWork > Movie

A movie.

Property	Expected Type	Description
Properties from Thing		
additionalType	URL	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.
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
sameAs	URL	URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Freebase page, or official website.
url	URL	URL of the item.
Properties from CreativeWork		
about	Thing	The subject matter of the content.
accountablePerson	Person	Specifies the Person that is legally accountable for the CreativeWork.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alternativeHeadline	Text	A secondary title of the CreativeWork.
associatedMedia	MediaObject	The media objects that encode this creative work. This property is a synonym for encodings.
audience	Audience	The intended audience of the item, i.e. the group for whom the item was created.
audio	AudioObject	An embedded audio object.
author	Organization or Person	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.
award	Text	An award won by this person or for this creative work.
awards	Text	Awards won by this person or for this creative work. (legacy spelling; see singular form, award)
citation	CreativeWork or Text	A citation or reference to another creative work, such as another publication, web page, scholarly article, etc. NOTE: Candidate for promotion to ScholarlyArticle.
comment	UserComments	Comments, typically from users, on this CreativeWork.
contentLocation	Place	The location of the content.
contentRating	Text	Official rating of a piece of content—for example,'MPAA PG-13'.

RDFa Lite

- Minimal subset of RDFa
- Specifically designed for schema.org type markup
- five simple attributes
  - vocab
  - typeof
  - property
  - resource
  - prefix

schema.org with RDFa Lite

```
<div vocab="http://schema.org" typeof="Movie">
  <h1 property="name">Avatar</h1>
  <span>Director: <span property="director">James Cameron</span>
(born August 16, 1954)</span>
  <span property="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" property="trailer">Trailer</a>
</div>
```

Advantages of RDFa Lite

<p vocab="http://schema.org/" **resource="#manu"** typeof="Person">
My name is
Manu Sporny
and you can give me a ring via .....

Universal identifier = http://some base address#manu

<p vocab="http://schema.org/" **prefix="ov: http://open.vocab.org/terms/"**
resource="#manu" typeof="Person">
My name is
Manu Sporny
and you can give me a ring via
1-800-555-0199.

My favorite animal is the <span **property="ov:preferredAnimal"**>Liger.
</p>

Microdata vs. RDFa Lite

Microdata 1.0	RDFa Lite 1.1	Purpose
itemid	resource	Used to identify the exact thing that is being described using a URL, such as a specific person, event, or place.
itemprop	property	Used to identify a property of the thing being described, such as a name, date, or location.
itemscope	not needed	Used to signal that a new thing is being described.
itemtype	typeof	Used to identify the type of thing being described, such as a person, event, or place.
itemref	not needed	Used to copy-paste a piece of data and associate it with multiple things.
not supported	vocab	Used to specify a default vocabulary that contains terms that are used by markup.
not supported	prefix	Used to mix different vocabularies in the same document, like ones provided by Facebook, Google, and open source projects.

Microdata 1.0

```
<div itemscope itemtype="http://schema.org/Product">
  <img itemprop="image" src="dell-30in-lcd.jpg" />
  <span itemprop="name">Dell UltraSharp 30" LCD Monitor</span>
</div>
```

RDFa Lite 1.1

```
<div vocab="http://schema.org/" typeof="Product">
  <img property="image" src="dell-30in-lcd.jpg" />
  <span property="name">Dell UltraSharp 30" LCD Monitor</span>
</div>
```

schema.org in JSON-LD

- Metadata not mixed with HTML elements
- Fully expressive like RDFa
- http://schema.org/Movie

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Movie",
  "actor": [
    {
      "@type": "Person",
      "name": "Johnny Depp"
    },
    {
      "@type": "Person",
      "name": "Penelope Cruz"
    },
    {
      "@type": "Person",
      "name": "Ian McShane"
    }
  ],
  "aggregateRating": {
    "@type": "AggregateRating",
    "bestRating": "10",
    "ratingCount": "200",
    "ratingValue": "8",
    "reviewCount": "50"
  },
  "author": [
    {
      "@type": "Person",
      "name": "Ted Elliott"
    },
    {
      "@type": "Person",
      "name": "Terry Rossio"
```