Novel approaches to metadata … improving the Web

Aarhus12/JBoye

Aarhus, Denmark 2012-11-08

Olle Olsson

World Wide Web Consortium (W3C)

Swedish Institute of Computer Science (SICS)

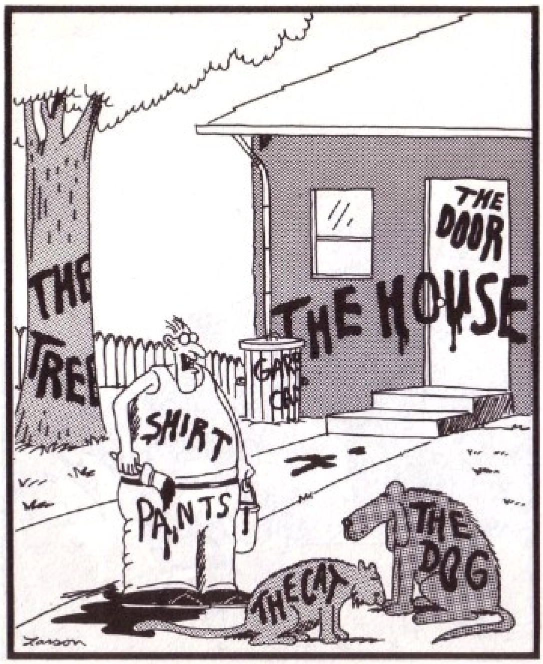
Contents

- What do we mean by metadata and semantics?
  - Enriched content for automated processing
- What are “light-weight” approaches?
  - Technologies
  - How they work
- Are these things really used?
  - Some statistics
- “What's in it for me?”
  - Examples
- So what is the conclusion?
  - !!!

What do we mean by metadata and semantics?

Web, metadata, semantics

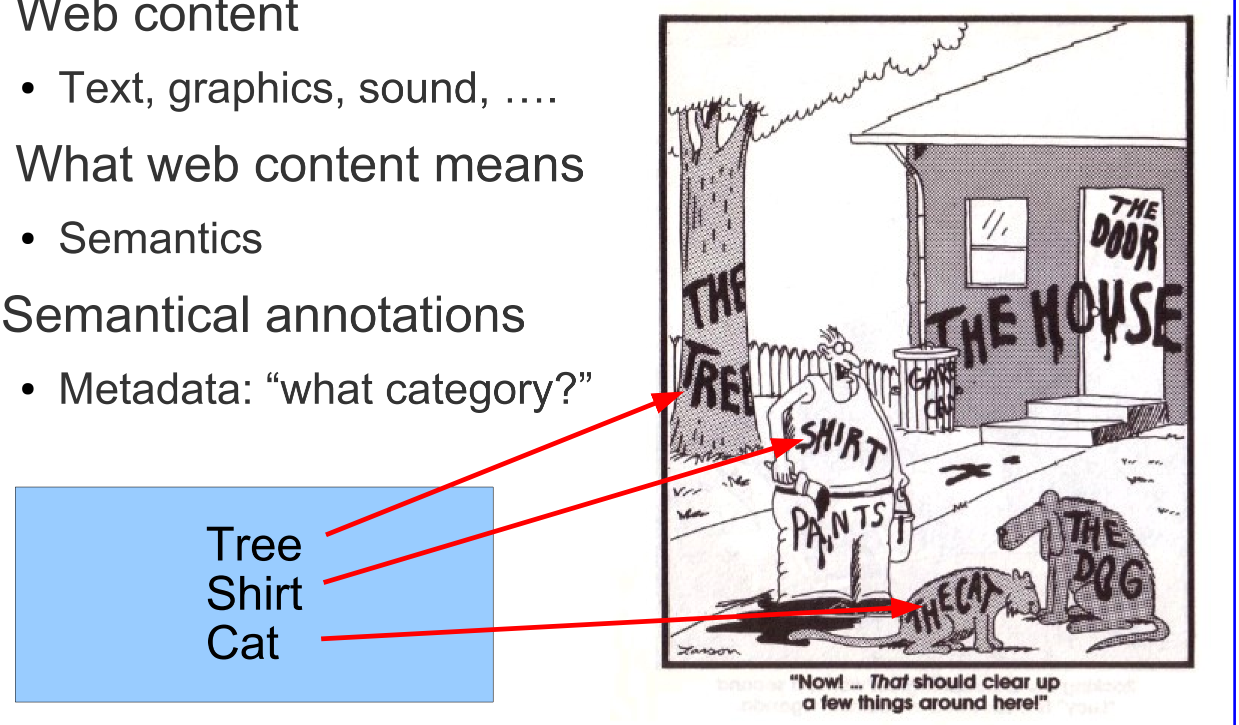
- Web content
  - Text, graphics, sound, ….
- What web content means
  - Semantics
- Semantical annotations
  - Metadata: “what category?”



"Now! ... That should clear up a few things around here!"

Web, metadata, semantics

- Web content
  - Text, graphics, sound, ….
- What web content means
  - Semantics
- Semantical annotations
  - Metadata: “what category?”



“Now! ... *That* should clear up a few things around here!”

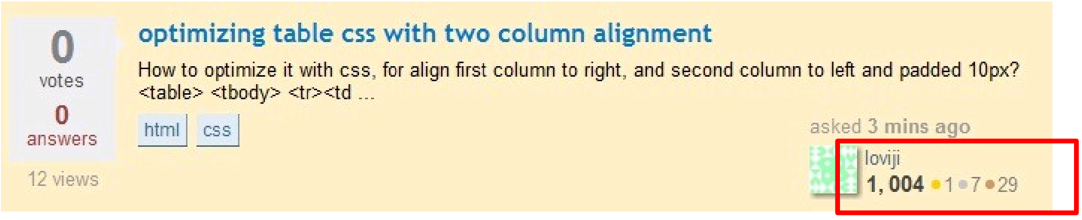
Purpose of Metadata / Semantics

- Be explicit about how to understand content
- Vocabulary
  - Terminology
  - Terms identify categories
- Not visible (not part of content)
- Used by mechanisms / tools / software
- Compare: what the web browser does
  - Terms identify structure and presentation
- Thesis: *A better web by metadata*

“Now! ... *That* should clear up a few things around here!”

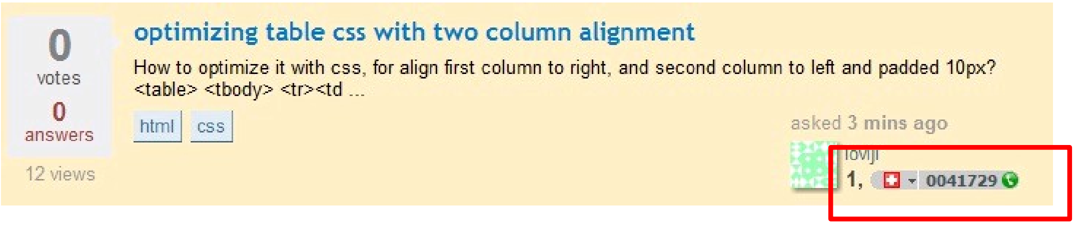
When not using semantic annotations

- Skype mistakes “reputation & badge count” for phone number



optimizing table css with two column alignment

How to optimize it with css, for align first column to right, and second column to left and padded 10px? <table> <tbody> <tr><td ...



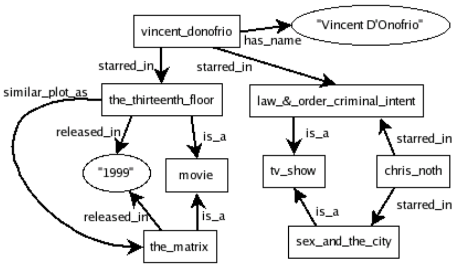
optimizing table css with two column alignment

How to optimize it with css, for align first column to right, and second column to left and padded 10px? <table> <tbody> <tr><td ...

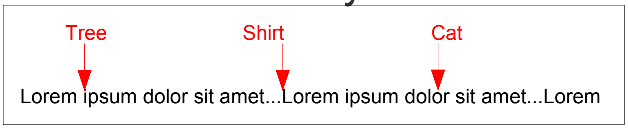
- Skype may misunderstand ISBN (international book number), numbers in general, etc
- Data not annotated … tools may make mistakes

“Big semantics” vs “Little semantics”

- Big semantics
  - Annotated data structure
  - At multiple levels



- Little semantics
  - Annotate key content parts
  - At elementary level



Lorem ipsum dolor sit amet...Lorem ipsum dolor sit amet...Lorem

What are "light-weight" approaches?

Light-weight approaches

- Adding extra markup inside HTML documents
  - HTML[5] / XML syntax
- Expresses category of a piece of content
  - Metadata
- Three used approaches:
  - Microformats
  - RDFa
  - Microdata

Microformats

- Created/defined by **`microformats.org`**
  - Independent, informal group
- Defines vocabularies for a set of specific needs
  - Terms indicate categories of human-readable content
- No new mark-up defined
  - Set values for attribute **`class`**

Microformats

- Example: hCard
  - “plain old HTML”

```
<p>
  The owner is
  <a href="http://facebook.com/John.Doe">John Doe</a>,
  who lives in
  London.
</p>
```

  - With microformat metadata

```
<p>
  The owner is
  <span class="vcard">
    <a class="url fn"
     href="http://facebook.com/John.Doe">John Doe</a>
  </span>,
  who lives in
  London.
</p>
```

Microformats

- Microformat: hCard
  - Based on vCard standard (RFC2426)
  - Used for people, organizations, contacts
- Other microformats
  - adr - address location information
  - geo - latitude & longitude location (WGS84 geographic coordinates)
  - hAtom - blog posts and other date-stamped content
  - hCalendar - events
  - hListing - listings for products or services
  - hMedia - media info about images, video, audio
  - hNews - news articles, extension of hAtom
  - hProduct - products
  - hResume - individual resumes and CVs
  - hReview - individual reviews and ratings

RDFa

- Created/defined by `w3c.org`
  - Web standard
- <u>No specific vocabularies</u> defined
  - General framework
- RDFa Lite 1.1 – extremely simplified
  - No new elements, but <u>new attributes</u>:
    - `vocab, typeof, property, resource, prefix`
- No constraint on what vocabularies to use
  - Many existing vocabularies

RDFa / Lite 1.1

- Example – Person
  - “plain old HTML”

```
<p>
  The owner is
  <a href="http://facebook.com/John.Doe">John Doe</a>,
  who lives in
  London.
</p>
```

  - With schema.org microdata metadata

```
<p>
  The owner is
  <span vocab="http://schema.org/" typeof="Person">
    <a property="url"
       href="http://facebook.com/John.Doe" >
     <span property="name">John Doe</span></a>
  </span>,
  who lives in
  London.
</p>
```

RDFa

- RDFa: part of the Semantic Web toolkit
  - “**RDF** in **a**ttributes”: use in HTML.
- Start small, go bigger when needed:
  - RDFa Lite 1.1 ← *Little semantics*
  - RDFa Core 1.1
  - HTML + RDFa 1.1
  - XHTML + RDFa 1.1
  - RDF
  - OWL ← *Big semantics*

Microdata

- Created/defined by **w3c.org**
  - Web standard
- No specific vocabularies defined
  - General framework
- No new elements, but new attributes:
  - **itemscope**, **itemtype**, **itemid**, **itemprop**, **itemref**
- No constraint on what vocabularies to use
  - Vocabularies defined elsewhere
- Example vocabulary
  - **schema.org**

Microdata/schema.org

- Example – Person
  - “plain old HTML”

```
<p>
  The owner is
  <a href="http://facebook.com/John.Doe">John Doe</a>,
  who lives in
  London.
</p>
```

- With schema.org microdata metadata

```
<p>
  The owner is
  <span itemscope itemtype="http://schema.org/Person">
    <a itemprop="url name"
       href="http://facebook.com/John.Doe">John Doe</a>
  </span>,
  who lives in
  London.
</p>
```

Microdata/schema.org

- Schema.org: defines 500+ categories/types, like:
  - `Person` - a person (alive, dead, undead, or fictional)
  - `Event` - an event happening at a certain time at a certain location
  - `Organization` - an organization such as a school, NGO, corporation, club, etc
  - `LocalBusiness` - a particular physical business or branch of an organization
  - `Place` - entities that have a fixed, physical extension.
  - `Product` - anything that is made available for sale

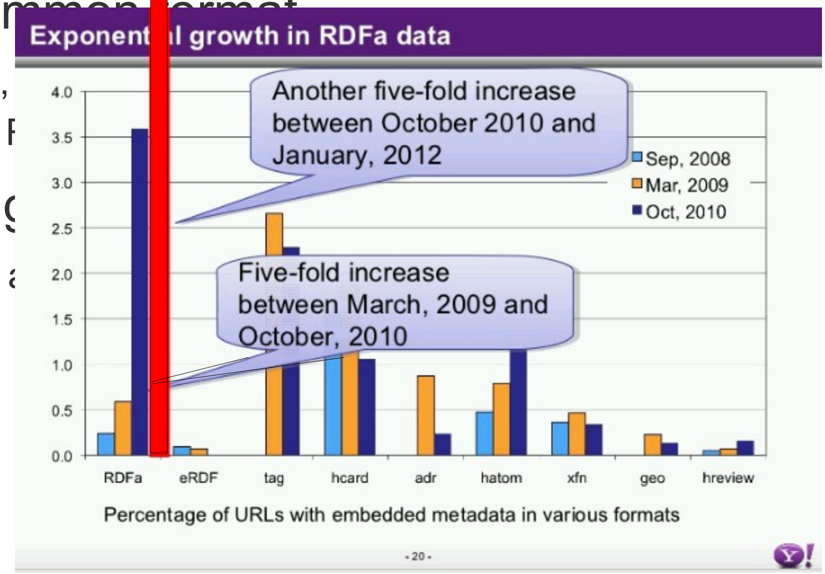
Are these things really used?

Statistics

- Crawl the web and inspect
- 31% of web pages, 5% of domains contain some metadata
  - Analysis of the Bing Crawl (US crawl, January 2012)
- RDFa is most common format
  - By URL: 25% RDFa, 7% microdata, 9% microformat
  - By eTLD (PLD): 4% RDFa, 0.3% microdata, 5.4% microformat
- Adoption is stronger among large publishers
  - Especially for RDFa and microdata

Statistics

- 31% of web pages, 5% of domains contain some metadata
  - Analysis of the Bing Crawl (US crawl, January 2012)
- RDFa is most co
  - By URL: 25% RDFa,
  - By eTLD (PLD): 4% R
- Adoption is strong
  - Especially for RDFa a



Exponential growth in RDFa data

Another five-fold increase between October 2010 and January, 2012

Five-fold increase between March, 2009 and October, 2010

Percentage of URLs with embedded metadata in various formats

Who? Top sites

RDFa

- Facebook.com
- Tabelog.com
- Venere.com
- Yahoo.com
- Tripadvisor.co.uk
- Answers.com
- Myspace.com
- Daodao.com
- Imdb.com
- Youtube.com
- Bestboy.com
- …..

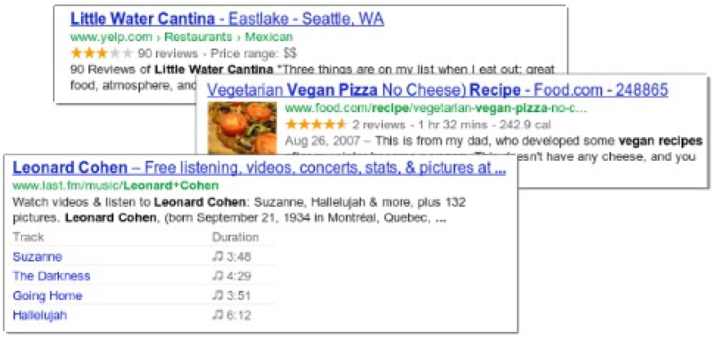
microdata

- Myspace.com
- Yelp.com
- Bbb.com
- Imdb.com
- Thefreelibrary.com
- Powells.com
- Youtube.com
- Homefinder.com
- Reverbnation.com
- Kino-teatr.ru
- Eventful.com
- …..

"What's in it for me?"

SEO

- Enhance web search
- Schema.org
  - Collaboration between: Google, Yahoo!, Microsoft/Bing
    - Started 2011
  - Standardised vocabulary for metadata mark-up
  - To be used by search engines
    - For better presentation of search results
    - For better precision in search
  - Google Rich Snippets



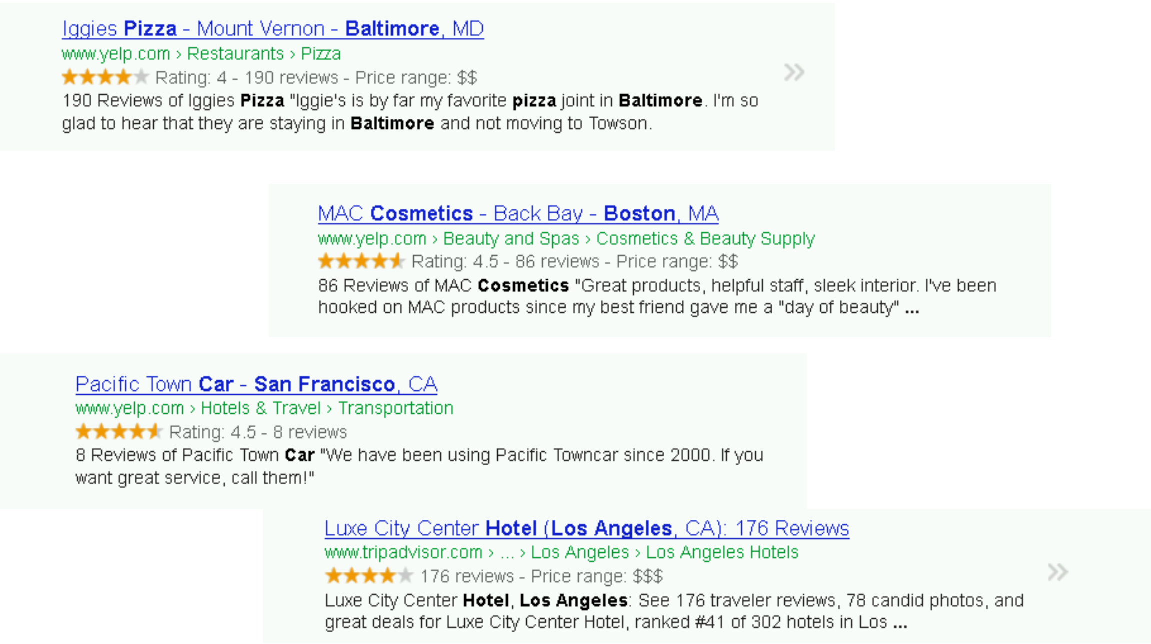
Little Water Cantina - Eastlake - Seattle, WA
www.yelp.com › Restaurants › Mexican
★★★☆☆ 90 reviews - Price range: $$
90 Reviews of **Little Water Cantina** "Three things are on my list when I eat out: great food, atmosphere, and

Vegetarian **Vegan Pizza** No Cheese) **Recipe** - Food.com - 248865
www.food.com/**recipe**/vegetarian-**vegan**-**pizza**-no-c...
★★★★★ 2 reviews - 1 hr 32 mins - 242.9 cal
Aug 26, 2007 – This is from my dad, who developed some **vegan recipes** ... esn't have any cheese, and you

Leonard Cohen – Free listening, videos, concerts, stats, & pictures at ...
www.last.fm/music/**Leonard+Cohen**
Watch videos & listen to **Leonard Cohen**: Suzanne, Hallelujah & more, plus 132 pictures. **Leonard Cohen**, (born September 21, 1934 in Montréal, Quebec, ...

Track	Duration
Suzanne	♫ 3:48
The Darkness	♫ 4:29
Going Home	♫ 3:51
Hallelujah	♫ 6:12

SEO: Example search results



Iggies **Pizza** - Mount Vernon - **Baltimore**, MD
www.yelp.com › Restaurants › Pizza
★★★★☆ Rating: 4 - 190 reviews - Price range: $$
190 Reviews of Iggies **Pizza** "Iggie's is by far my favorite **pizza** joint in **Baltimore**. I'm so glad to hear that they are staying in **Baltimore** and not moving to Towson.

MAC **Cosmetics** - Back Bay - **Boston**, MA
www.yelp.com › Beauty and Spas › Cosmetics & Beauty Supply
★★★★★ Rating: 4.5 - 86 reviews - Price range: $$
86 Reviews of MAC **Cosmetics** "Great products, helpful staff, sleek interior. I've been hooked on MAC products since my best friend gave me a "day of beauty" ...

Pacific Town **Car** - **San Francisco**, CA
www.yelp.com › Hotels & Travel › Transportation
★★★★★ Rating: 4.5 - 8 reviews
8 Reviews of Pacific Town **Car** "We have been using Pacific Towncar since 2000. If you want great service, call them!"

Luxe City Center **Hotel** (**Los Angeles**, CA): 176 Reviews
www.tripadvisor.com › ... › Los Angeles › Los Angeles Hotels
★★★★☆ 176 reviews - Price range: $$$
Luxe City Center **Hotel**, **Los Angeles**: See 176 traveler reviews, 78 candid photos, and great deals for Luxe City Center Hotel, ranked #41 of 302 hotels in Los ...

SEO: syntax vs semantics

Note about enriched search results:

- Search engines can handle all light-weight formats
- They do prefer rich vocabularies for semantic annotation
- Vocabulary: `Schema.org` works equally well in:
  - `RDFa` syntax
  - `Microdata` syntax

More benefits

- Web apps:
  - discover data about your website; use them to interface with data on your site.
- Browser extensions:
  - offer new user actions
    - copy contact to address book
    - add event to calendar
    - present geolocation on map
- Aggregators
  - collect relevant data from your page

So what is the conclusion?

- New light-weight metadata approaches
  - Technically simple to use
    - Just extend the templates for generating ordinary page contents
  - Brings direct advantages to SEO
- Richer web content
  - Enables more reuse / re-purposing of data
  - Enables more automatic processing of contents
  - More processing in the client: need data, not only strings of letters
- Strong support from major players
  - Schema.org pushed strongly by search engine companies

Thank you for your attention!