Semantic Web and Microformats

Web Architecture and Information Management [./] Spring 2009 — INFO 190-02 (CCN 42509)

Erik Wilde, UC Berkeley School of Information

2009-04-27

Contents

Abstract (2)

HTML pages are for human users and describe a resource in structural terms (headings, lists, tables, …). For machine-based interaction, it is often useful to have more information about the application concepts. The Web reflects the various ways in which the issue of *semantics* has been addressed in other disciplines, with the *Semantic Web* having the strongest commitment to highly formalized semantics. On the syntax side, the *Extensible Markup Language (XML)* is a popular language for representing application structures, but it is representing only syntax and no semantics

HTML vs. XML (3)

- HTML describes structures in a very general way
  - HTML elements describe logical page structures such as headings, lists, tables, …
  - useful for dynamic and adaptive page rendering, but not for understanding contents
- Good HTML may have more information available
  - classes in HTML elements may represent underlying concepts (CSS may use this)
  - HTML containers [Advanced HTML; All-Purpose Elements (1)] may represent aggregation of some basic information items
- Very good HTML
  - some guidelines/rules/methods for *understanding* class names
  - some model for the underlying schema (what may appear in which combination)
- Excellent HTML is dynamically generated from XML
  - the model is exposed as structured XML data that is available to the client
  - there is a stylesheet for producing the HTML version of the XML
  - but even XML does not provide semantics (it is just a structured syntax)

Plain HTML (4)

```
<html>
   <head>
      <title>Cannondale 2007 System Six 2</title>
   </head>
   <body>
      <h1>Cannondale 2007 System Six 2</h1>
      <ul>
         <li>Mavic Ksyrium ES wheelset</li>
         <li>Maxxis Xenith Hors Categorie tires</li>
         <li>Fi'zi:k Arione Titanium saddle</li>
         <li>SRAM Force components</li>
      </ul>
      <p>Sizes: 48cm, 50cm, 52cm, 54cm, 56cm, 58cm, 60cm, 63cm</p>
      <p>Dealers: </p>
      <ul>
         <li>Mike's Bikes of Berkeley, 2161 University Avenue, Berkeley, CA
94704; +1-510-8452453</li>
      </ul>
   </body>
</html>
```


Good HTML (5)

```
<html>
   <head>
      <title>Cannondale 2007 System Six 2</title>
   </head>
   <body>
      <h1 class="bike"><span class="manufacturer">Cannondale</span> <span
class="year">2007</span> <span class="model">System Six</span> <span
class="type">2</span></h1>
      <ul class="components">
         <li class="component"><span class="manufacturer">Mavic</span> <span
class="type">Ksyrium ES</span> wheelset</li>
         <li class="component"><span class="manufacturer">Maxxis</span> <span
class="type">Xenith Hors Categorie</span> tires</li>
         <li class="component"><span class="manufacturer">Fi'zi:k</span>
<span class="type">Arione Titanium</span> saddle</li>
         <li class="component"><span class="manufacturer">SRAM</span> <span
class="type">Force</span> components</li>
      </ul>
      <p>Sizes: <span class="size">48cm</span>, <span
class="size">50cm</span>, <span class="size">52cm</span>, <span
class="size">54cm</span>, <span class="size">56cm</span>, <span
class="size">58cm</span>, <span class="size">60cm</span>, <span
class="size">63cm</span></p>
      <p>Dealers: </p>
      <ul>
         <li class="dealer"><span class="name">Mike's Bikes of
Berkeley</span>, <span class="adr"><span class="street-address">2161
University Avenue</span>, <span class="locality">Berkeley</span>, <span
class="region">CA</span> <span class="postal-code">94704</span></span>; <a
href="tel:+1-510-8452453">+1-510-8452453</a></li>
      </ul>
   </body>
</html>
```

Excellent HTML (6)

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet href="bike2html.xsl" type="text/xsl"?>
<bike manufacturer="Cannondale" year="2007">
 <model>System Six</model>
 <type>2</type>
 <sizes>
  <size unit="cm">48</size>
  <size unit="cm">50</size>
  <size unit="cm">52</size>
  <size unit="cm">54</size>
  <size unit="cm">56</size>
  <size unit="cm">58</size>
  <size unit="cm">60</size>
  <size unit="cm">63</size>
 </sizes>
 <parts>
  <wheelset manufacturer="Mavic">Ksyrium ES</wheelset>
  <tires manufacturer="Maxxis">Xenith Hors Categorie</tires>
  <saddle manufacturer="Fi'zi:k">Arione Titanium</saddle>
  <components manufacturer="SRAM">Force</components>
 </parts>
 <dealers>
  <dealer>
   <name>Mike's Bikes of Berkeley</name>
   <address>2161 University Avenue</address>
   <city>Berkeley</city>
   <zip>94704</zip>
   <state>CA</state>
   <phone>+1-510-8452453</phone>
  </dealer>
 </dealers>
</bike>
```


XML → HTML Stylesheet (7)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL
/Transform">
 <xsl:template match="/">
  <html>
   <head>
    <title>
     <xsl:value-of select="bike/@manufacturer"/>
     <xsl:text> </xsl:text>
     <xsl:value-of select="bike/@year"/>
     <xsl:text> </xsl:text>
     <xsl:value-of select="bike/model"/>
     <xsl:text> </xsl:text>
     <xsl:value-of select="bike/type"/>
    </title>
   </head>
   <body>
    <h1 class="bike">
     <span class="manufacturer">
      <xsl:value-of select="bike/@manufacturer"/>
     </span>
     <xsl:text> </xsl:text>
     <span class="year">
      <xsl:value-of select="bike/@year"/>
     </span>
     <xsl:text> </xsl:text>
     <span class="model">
      <xsl:value-of select="bike/model"/>
     </span>
     <xsl:text> </xsl:text>
     <span class="type">
      <xsl:value-of select="bike/type"/>
     </span>
    </h1>
    <ul class="components">
     <xsl:for-each select="//parts/*">
      <li class="component">
       <span class="manufacturer">
        <xsl:value-of select="@manufacturer"/>
       </span>
       <xsl:text> </xsl:text>
       <span class="type">
        <xsl:value-of select="text()"/>
```

Graceful Degradation (8)

- XML was designed as a language for Web content
  - the idea was that XML documents would be delivered to the browser
  - stylesheets (CSS/XSL) would take care of the client-side rendering
- CSS [Cascading Style Sheets (CSS)] is good at supporting graceful degradation
  - viewing an HTML page with CSS turned off most of the time works fine
- XSL is not good at supporting graceful degradation
  - the browser just displays the raw XML when XSL is not supported
- Serving XML on the Web is not a good idea
  - in closed scenarios (intranet applications) this might be a viable solution
  - in open scenarios, HTML should be served as the default representation
  - alternate versions can be provided by supporting *HTTP content Negotiation*

Excellent HTML (9)

```
<html>
   <head>
      <title>Cannondale 2007 System Six 2</title>
      <link title="XML version" rel="alternate" type="application/xml"
href="systemsix.xml"/>
   </head>
   <body>
      <h1 class="bike"><span class="manufacturer">Cannondale</span> <span
class="year">2007</span> <span class="model">System Six</span> <span
class="type">2</span></h1>
      <ul class="components">
         <li class="component"><span class="manufacturer">Mavic</span> <span
class="type">Ksyrium ES</span> wheelset</li>
         <li class="component"><span class="manufacturer">Maxxis</span> <span
class="type">Xenith Hors Categorie</span> tires</li>
         <li class="component"><span class="manufacturer">Fi'zi:k</span>
<span class="type">Arione Titanium</span> saddle</li>
         <li class="component"><span class="manufacturer">SRAM</span> <span
class="type">Force</span> components</li>
      </ul>
      <p>Sizes: <span class="size">48cm</span>, <span
class="size">50cm</span>, <span class="size">52cm</span>, <span
class="size">54cm</span>, <span class="size">56cm</span>, <span
class="size">58cm</span>, <span class="size">60cm</span>, <span
class="size">63cm</span></p>
      <p>Dealers: </p>
      <ul>
         <li class="dealer"><span class="name">Mike's Bikes of
Berkeley</span>, <span class="adr"><span class="street-address">2161
University Avenue</span>, <span class="locality">Berkeley</span>, <span
class="region">CA</span> <span class="postal-code">94704</span></span>; <a
href="tel:+1-510-8452453">+1-510-8452453</a></li>
      </ul>
   </body>
</html>
```

From Information, Knowledge (10)

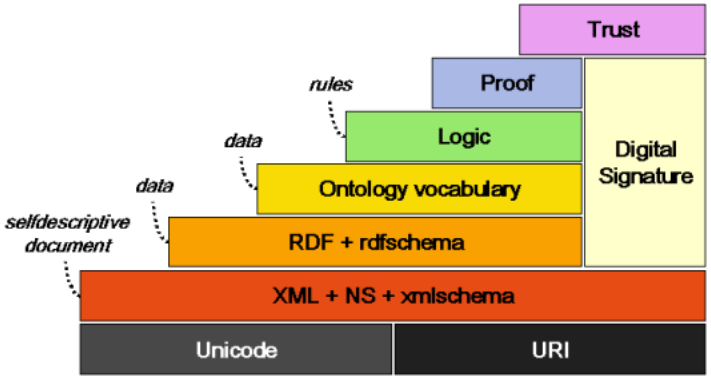
- XML is often said to be "self-describing"
  - many people think this is the same as "self-explanatory"
  - the catch is what exactly it is you refer to by "describing"
- Database data cannot live without a database
  - database data is simply content, the structure is provided by a DBMS
  - XML documents have their structure encoded within them
  - compared to database data, XML in fact is "self-describing"
- What is the gap between "self-describing" and "self-explanatory"?
  - it is impossible to find out how the document could be modified
  - there are no semantics associated with structure or content
  - so "self-describing" means, you can guess a lot, but you maybe wrong

The Semantic Web Hype (11)

1965, H. A. Simon: "machines will be capable, within twenty years, of doing any work a man can do" [http://en.wikipedia.org/wiki/Artificial_intelligence#_note-11]
1967, Marvin Minsky: "Within a generation […] the problem of creating 'artificial intelligence' will substantially be solved." [http://en.wikipedia.org/wiki/Artificial_intelligence#_note-12]

- How to get past the limitations of HTML?
  - a machine-friendly Web must make Web resources machine-processable
  - XML solved the problem on the syntax level
  - how could the problem be solved on the level of semantics?
- As in the 1970's, *description logic* was declared as being the solution
  - there was a need for the Web to move towards semantics
  - there was a community of AI researchers with a long history
  - the *Semantic Web* was born and is currently repeating AI history

Semantic Web Layers (12)



Conclusions (13)

- Formal semantics allow machines to "understand" Web content
- Semantics can be "deep" or "shallow"
- Philosophy has discussed these question for a log time
- The Web has two major approaches for expressing semantics:
  - *Semantic Web* as an attempt to completely formalize semantics
  - *Microformats* as a pragmatic way of agreeing on useful bits and pieces