

Hypertext Markup Language (HTML)

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

Erik Wilde, UC Berkeley School of

Information

2009-01-28



<http://creativecommons.org/licenses/by/3.0/>

This work is licensed under a CC

[Attribution 3.0 Unported License](http://creativecommons.org/licenses/by/3.0/) [http://creativecommons.org/licenses/by/3.0/]

Contents

• Abstract	2
• Structured Documents on the Web	3
• 1 HTML Validation	
◦ Checking for Correctness	5
◦ Tool-Based Validation	6
◦ Web-Based Validation	7
◦ Firebug	8
• HTML and WYSIWYG	9
• Web Browsers	10
• 2 HTML and Structure	
◦ Text	12
◦ More Advanced Text	13
◦ Lists	14
◦ Tables	15
◦ Merging Table Cells	16
◦ Merging Rows	17
◦ Merging Columns	18
◦ Images	19
◦ Image Details	20
◦ Links	21
• 3 Conclusions	
◦ HTML Matters	23

Abstract (2)

The *Hypertext Markup Language (HTML)* is the most important content type on the Web. This lecture covers a basic overview of how to use HTML markup in general. In particular, we look at page titles, meta tags, inserting text and images, using lists, and creating simple tables. Attributes can be used for more layout control in the HTML tags, but most layout issues are deferred until the CSS lecture.

Structured Documents on the Web (3)

- *Hypertext Markup Language (HTML)* is the language of the Web
 - *Hypertext* because the Web is a hypermedia system
 - *Markup* because documents are encoded using text
 - *Language* because HTML is used for communications
- *Markup Languages* are different from most file formats
 - many computer formats are binary encoded and not “just text”
 - *markup* allows structured documents to be encoded *as just text*
- Web data formats use markup as well as other encodings
 - *HTML* and *XML* are markup languages
 - *JavaScript* is also exchanged textually (but it's not markup)
 - images and other multimedia content is encoded as binary files

HTML Validation

Checking for Correctness (5)

- HTML's structure is defined by a *Document Type Definition (DTD)*
 - formally speaking, a DTD defines the grammar of the HTML language
 - (and if you really want to know, *SGML* defines the syntax)
 - colloquially speaking, a DTD defines how to combine elements and attributes

```
<!-- Unordered Lists (UL) bullet styles -->
<!ELEMENT UL - - (LI)+          -- unordered list -->
<!ATTLIST UL
  %attrs;                       -- %coreattrs, %i18n, %events
--
  >

<!ELEMENT LI - O (%Flow;)*      -- list item -->
<!ATTLIST LI
  %attrs;                       -- %coreattrs, %i18n, %events
--
  >

<!ENTITY % flow "%block; | %inline;">

<!ENTITY % block
  "P | %heading; | %list; | %preformatted; | DL | DIV | NOSCRIPT |
  BLOCKQUOTE | FORM | HR | TABLE | FIELDSET | ADDRESS">
```

Tool-Based Validation (6)

- Testing HTML makes sure that it is well-defined
 - any errors have to be corrected by the browser
 - the results of browser-side corrections are hard to predict
- HTML editors allow validation within the tool
 - in theory, using the [public DTD](http://www.w3.org/TR/html4/loose.dtd) [http://www.w3.org/TR/html4/loose.dtd], in practice, using a [local copy](#) [src/html4-loose.dtd]
- In theory, HTML editors should always produce valid HTML
 - things today are not as bad as they used to be
 - creating valid HTML can be a challenge for complex Web pages

Web-Based Validation (7)

- [Tool-Based Validation](#) [Tool-Based Validation (1)] requires locally installed tools
 - maybe hard to install and hard to maintain across computers
 - for power users, locally installed tools are hard to beat
- *Web-based tools* allow validation from anywhere
- [W3C](http://www.w3.org/) [http://www.w3.org]'s [markup validation service](http://validator.w3.org/) [http://validator.w3.org/] supports three modes:
 1. validation by URI (pointing at a random Web page)
 2. validation by file upload (allows validation of non-Web files)
 3. validation by copy/paste (lightweight mode for small experiments)
- Markup validation is only one facet of checking Web content
 - [checking CSS code for validity](http://jigsaw.w3.org/css-validator/) [http://jigsaw.w3.org/css-validator/]
 - [checking Web pages for mobile content \(i.e., simpler HTML\)](http://validator.w3.org/mobile/) [http://validator.w3.org/mobile/]
 - [checking Web pages for broken links](http://validator.w3.org/checklink) [http://validator.w3.org/checklink]

Firebug (8)

- Browser-based inspection (much better than "view source")
- Learning Web design by looking at Web design
 - Firefox's *View* → *Page Source* provides access to a page's source
 - Firefox's *Tools* → *Page Info* provides access to all ancillary files
- Understanding how complex HTML works is hard
 - looking at the source requires "brain-based rendering"
 - looking at a rendered document makes it hard to see the source
 - Firebug provides a convenient inspection feature for Web pages
- Inspection allows both directions of understanding HTML
 - inspecting the rendered page and looking at the source part
 - inspecting the source and seeing how it is being rendered
- Firebug also displays the associated [CSS](#) [Cascading Style Sheets (CSS)] code

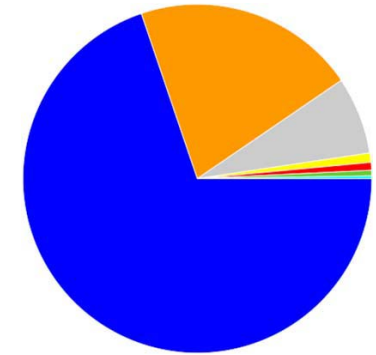


HTML and WYSIWYG (9)

- Thinking of HTML as a page-description language is wrong
 - the world's worst Web page editor: [Yahoo! PageBuilder](http://geocities.yahoo.com/v/pb.html)
[<http://geocities.yahoo.com/v/pb.html>]
- HTML has been designed as a structure-description language
 - structured contents can be reformatted and reflowed
- HTML rendering depends on a many client properties
 - screen/window size, resolution, and color depth
 - different available fonts
 - fonts with the same family name but different metrics
 - different hyphenation algorithms
 - hyphenation setting defaults
 - hyphenation dictionaries
 - different size spaces
 - different line-breaking algorithms
 - different widow/orphan/kepttogether rules

Web Browsers (10)

- Internet Explorer (69.80%)
- Mozilla Firefox (20.66%)
- Safari (7.18%)
- Chrome (0.87%)
- Opera (0.72%)
- Netscape (0.52%)
- Other (0.25%)



HTML and Structure

Text (12)

- h1-h6 are different levels of [headings](http://www.w3.org/TR/REC-html40/struct/global.html#h-7.5.5) [http://www.w3.org/TR/REC-html40/struct/global.html#h-7.5.5]
- p contains [paragraph text](http://www.w3.org/TR/REC-html40/struct/text.html#h-9.3.1) [http://www.w3.org/TR/REC-html40/struct/text.html#h-9.3.1]
 - whitespace and line wrapping are ignored
 - paragraphs are set as boxes containing a number of lines
- Text inside paragraphs can use additional markup ("[phrase markup](http://www.w3.org/TR/REC-html40/struct/text.html#h-9.2.1)") [http://www.w3.org/TR/REC-html40/struct/text.html#h-9.2.1]
 - em for *emphasized text*
 - strong for text with a **strong emphasis**
 - sub for subscript text
 - sup for superscript text
 - q for "quoted text" ("try 'nesting' quotes")
 - code for code examples
- rendering of all these elements is built into the browser
 - more sophisticated issues probably [are more browser-dependent](http://dret.typepad.com/dretblog/2008/04/internationaliz.html) [http://dret.typepad.com/dretblog/2008/04/internationaliz.html]

More Advanced Text (13)

- Quotations can be explicitly marked up as such
 - blockquote for block-level quotations
 - q for inline quotations (part of a block)
 - cite provides support for "pointing to the source"
- Preformatted text allows text formatting in the HTML source
 - pre leaves whitespace intact and usually uses monospaced fonts
 - word wrapping may be turned off by default

```
<br/>
<pre>This is some text that will not be wrapped because of
pre's default behavior,
at least in most browsers.</pre>
<br/>
<pre class="wrap">This is some text that will be wrapped
because of pre's overridden behavior,
at least in most browsers.</pre>
<br/>
```

Lists

(14)

- HTML supports three kinds of lists
 - ul for [unordered lists](http://www.w3.org/TR/REC-html40/struct/lists.html#h-10.2) containing li
 - ol for [ordered lists](http://www.w3.org/TR/REC-html40/struct/lists.html#h-10.2) containing li
 - dl for [definition lists](http://www.w3.org/TR/REC-html40/struct/lists.html#h-10.3) containing dt/dd

```
<d1>
<dt>Unordered Lists</dt>
<dd>
  <ul>
    <li>Unordered lists contain just lists of items</li>
    <li>Itemization symbols are chosen by the browser</li>
  </ul>
</dd>
<dt>Ordered Lists</dt>
<dd>
  <ol>
    <li>Ordered lists contain ordered lists of items</li>
    <li>The numbering scheme is chosen by the browser
      <ol>
        <li>Lists may be nested as deeply as required</li>
      </ol>
    </li>
  </ol>
</dd>
</d1>
```

Tables

(15)

- Tables are the most complex visual structure in HTML
 - [table](http://www.w3.org/TR/REC-html40/struct/tables.html#h-11.2.1) represents a table as a sequence of rows
 - [tr](http://www.w3.org/TR/REC-html40/struct/tables.html#h-11.2.5) represents a *table row* as a sequence of cells
 - [td](http://www.w3.org/TR/REC-html40/struct/tables.html#h-11.2.6) represents a table cell containing *table data*
 - [th](http://www.w3.org/TR/REC-html40/struct/tables.html#h-11.2.6) is a special cell containing *header data*

```
<table>
<tr>
  <th>Date</th>
  <th>Topic</th>
</tr>
<tr>
  <td>2009-01-28</td>
  <td><a href="../html-basic">HTML Basics</a></td>
</tr>
<tr>
  <td>2009-02-02</td>
  <td><a href="../html-advanced">Advanced HTML</a></td>
</tr>
</table>
```

Merging Table Cells (16)

A test table with merged cells

	Average		Red eyes
	height	weight	
Males	1.9	0.003	40%
Females	1.7	0.002	43%

Merging Rows (17)

- Table cells can span more than table row
 - rowspan specifies how many rows a cell is spanning
 - subsequent table rows must leave that space "empty"

```
<table>
<tr>
  <th>Date</th>
  <th>Topic</th>
</tr>
<tr>
  <td>2009-03-23</td>
  <td rowspan="2">Spring<br>Break</td>
</tr>
<tr>
  <td>2009-03-25</td>
</tr>
</table>
```


Merging Columns (18)

- Table cells can span more than table column
 - colspan specifies how many columns a cell is spanning
 - following cells in the same row must be left "empty"

```
<table>
<tr>
  <th>Date</th>
  <th colspan="2">Resources</th>
</tr>
<tr>
  <td>2009-01-28</td>
  <td><a href=" ../html-basic">HTML Basics</a></td>
  <td><a href=" ../2009-01-28-html-basic.pdf">PDF</a></td>
</tr>
<tr>
  <td>2009-02-02</td>
  <td><a href=" ../html-advanced">Advanced HTML</a></td>
  <td><a href=" ../2009-02-02-html-advanced.pdf">PDF</a></td>
</tr>
```

Images (19)

- The Web is an open hypermedia system
 - *hyper* refers to the term "hypertext" for linked content
 - *media* refers to the fact that multiple media types are supported
- For a long time, the Web only supported text and images
 - images can be used in a variety of formats (GIF, JPEG, PNG)
 - audio and video are possible today, but not "part of the Web"
- Images are not part of a Web page, they are included by markup
 - `img` [<http://www.w3.org/TR/REC-html40/struct/objects.html#h-13.2>] is an empty element for including images
 - `src` is a URI pointing to the image (often relative)

```

```

Image Details (20)

- Images must use a format supported by the browser
 - GIF, JPEG, and PNG are pretty safe choices
 - HTTP allows the browser to understand the image format
 - limited browsers might have size/complexity restrictions
- Image information increases accessibility of a page
 - alt contains a short description of the image
 - for icons it is essential to provide this information
- Image dimensions and image rendering
 - width/height specify the dimensions of the image
 - allows the browser to start rendering the page before the images are received
 - will be used to resize the image if the real image size is different
 - browser-based scaling of images [often is not a good idea](http://offog.org/articles/image-scaling/) [http://offog.org/articles/image-scaling/]

Links (21)

- Links are the most important feature of the Web
 - conceptually, the Web is one large hypermedia document
 - links are based on Web identifiers, the *Uniform Resource Identifier (URI)*
- a is a link *anchor* and links to a URI (the *link target*)

```
<a href="http://www.berkeley.edu" title="UC Berkeley">UCB</a>
```
- URIs can have various forms
 - http: points to resources available on Web servers
 - https: is the same but uses encrypted connections
 - URIs can use a variety of other *URI schemes*
 - URIs can be relative (in the same way as file names)
 - relative URIs are evaluated relative to the URI of their occurrence
 - relative URIs can use path segments such as "/" and "."

Conclusions

HTML Matters (23)

- HTML is not just getting text displayed
- Good HTML allows better browsing
- First represent as much as possible in HTML
- Then add what is missing as [CSS](#) [Cascading Style Sheets (CSS)] and/or microformats
- Graceful degradation is important