Session 3

HTML

Tim Berners-Lee

Reading

- HTML tutorials
  - www.w3schools.com/html/
  - en.wikipedia.org/wiki/Html

- Character sets
  - en.wikipedia.org/wiki/Character_encodings_in_HTML
Session 3 – HTML

References

- WWW Consortium - HTML 5
  https://www.w3.org/TR/html5/
- HTML character entity references
  www.htmlhelp.com/reference/html40/entities/

Lecture Objectives

- Become familiar with HTML syntax
- Understand the relationship between an HTML document and the corresponding element tree
- Understand the evolution of HTML
Evolution of HTML

- Began as a subset of SGML
- Implemented as vendor standards
- Evolved to vendor independent standards that were well implemented by vendors
- Continued evolution to remove styling
- Further evolved to XML structure
- HTML5 is the current standard, reducing burden of SGML legacy (and XML)

Example Document Structure

- An HTML page is a tree of html elements

An element (e.g., body) may contain other elements and/or text
HTML Element

An element consists of a begin tag and an end tag

```html
<td><div align="center">
    <img src="original_files/contact-us.gif" width="126" alt="" height="51" />
</div>
Fill out the form below and your information will be sent to a sales representative.
Be sure to specify what price range you are wanting to stay in.
If you prefer, you can call us toll free at 877-456-7223 or </td>
```

**Contact Us**

Fill out the form below and your information will be sent to a sales representative. Be sure to specify what price range you are wanting to stay in. If you prefer, you can call us toll free at 877-456-7223 or

---

What are the Components of HTML?

- Tag
- Element name

```html
<html>
<body bgcolor="#ffffff" ...

An element name appears in a start tag and (usually) in an end tag

<body> and </BODY>
```

Must be properly nested
An end tag closes all intervening tags
Element names are case insensitive in html
What are the Components of HTML?

`<body bgcolor="#FFFFFF" ...`

Attribute values can be set by authors, scripts, or by default

Attribute name/value pairs are separated by spaces

Attribute name/value pairs may appear in any order

Attribute Values

- Are usually enclosed in quotes (single or double), but quotes are not required in HTML if the value of the attribute does not contain special characters
- Always required in XHTML
- May be restricted to a specific set of values

Important when you are enclosing HTML in a programming language String
Terms to Know

- **Document** - a message entity with a content type of text/html (also applies to other text documents)
- **HTML user agent** - a device that interprets HTML documents (includes browsers)

HTML Generation

- Most HTML is generated by some WYSIWYG tool
- Usually the generated HTML has some deviation from pure HTML syntax
- Imperfect HTML is usually not a problem since browsers have become better in handling of HTML syntax errors
Doctype

<!DOCTYPE html >

- First line of your document
- Identifies the version of HTML the document should comply with
- Reference to DTD no longer needed (with HTML5)
- Above example will be validated as HTML5

Head Element

- The head element contains header information about the document, such as its title, keywords, description, and style sheet

```html
<head>
<title>2008 CSE336 Conference</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<link href="Registration_files/dt_javaone.css" rel="stylesheet" type="text/css" />
<style type="text/css">
.smaller {
  font-size:11px;
  TEXT-DECORATION: none
}
</style></head>
```
Meta Tag

- Metadata - information about data.
- `<meta>` tag provides non-displayed metadata about the document.
- Metadata can be used by browsers, search engines, or other web services.
- HTML5 introduced a meta element to let web designers take control over the viewport (the user’s visible area of a web page).

Body Elements

- Viewable content in HTML (e.g., text, images, graphics) is painted (top to bottom) onto the visible page for visual browsers.
- Contains elements that are either block level or inline:
  - Block level - begin on a new line
  - Inline - text level
  - div and span are used to provide additional structure (block and inline)
Text Elements

<body>
  <p>This is a paragraph tag</p>
  <ul>
    <li>This is an item in an unordered list.</li>
    <li>This is another item in that list.</li>
  </ul>
  <ol>
    <li>This is an item in an ordered list.</li>
    <li>This is another item in that list.</li>
  </ol>
</body>

Definition lists are also available in HTML

Text

- Inline elements - em, strong, cite, code, abbr, acronym, Q, sub, sup, etc.
- Block elements - blockquote, p

The elements that dictate appearance are best replaced by CSS (covered in the next session)
Characters

```html
<meta charset="UTF-8">
```

- The meta element can be used to communicate communications protocol information to the server.
- You should place the information early in the document head element.

Document Representations

- Servers send HTML documents to agents as a bytestream; user agents interpret them as a sequence of characters.
- HTML allows different computers to interoperate seamlessly, but these computers may use different character encodings.
- This process requires a knowledge of:
  - Document character set - characters used in a document.
  - Character encodings - the byte representations of characters - referred to as “charset.”
Early Character Codes

- The earliest encoding systems used six bits (BCD), allowing 64 characters.
- In 1963:
  - 8-bit EBCDIC was introduced by IBM.
  - The 7-bit ASCII code was introduced and used by other computer HW manufacturers.
- The codes are:
  - Clearly inadequate for global commerce.
  - Important to understand implementation of current codes (backwards compatibility).

Characters

- Languages consist of a set of characters, usually defined as the smallest unit of information in the written form of a natural language.
- Examples:
  - English includes 26 letters (a-z), along with their capital equivalents, digits (0-9), and special symbols (e.g., ,,.)
  - Chinese has 4,000 characters for general language coverage and 40,000 characters for more complete coverage.
  - Japanese has 2,000 characters for general language coverage.
  - There are approximately 6,800 living languages in the world today.
Character Code Issues

- Character codes
  - Mapping of characters to strings of binary digits
  - E.g., “S” usually is usually mapped to 01000011₂
- Mapping to an 8-bit code usually restricts the language to 256 characters
- Mapping to longer character codes can result in longer strings
  - Length of text strings still a concern, even with much less expensive memory and disk
  - Text is sometimes transmitted over low bandwidth communications links
    Each mapping is sometimes referred to as a “code point”

ASCII Reference Table

Note the ordering of characters

<table>
<thead>
<tr>
<th>MSD</th>
<th>LSD</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NUL</td>
<td>DLE</td>
<td>SP</td>
<td>0</td>
<td>@</td>
<td>P</td>
<td>p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>SOH</td>
<td>DC1</td>
<td>!</td>
<td>1</td>
<td>A</td>
<td>Q</td>
<td>a</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>STX</td>
<td>DC2</td>
<td>*</td>
<td>2</td>
<td>B</td>
<td>R</td>
<td>b</td>
<td>r</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ETX</td>
<td>DC3</td>
<td>#</td>
<td>3</td>
<td>C</td>
<td>S</td>
<td>c</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>EOT</td>
<td>DC4</td>
<td>$</td>
<td>4</td>
<td>D</td>
<td>T</td>
<td>d</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ENQ</td>
<td>NAK</td>
<td>%</td>
<td>5</td>
<td>E</td>
<td>U</td>
<td>e</td>
<td>u</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>AC1</td>
<td>SYN</td>
<td>&amp;</td>
<td>6</td>
<td>F</td>
<td>V</td>
<td>f</td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>BEL</td>
<td>ECH</td>
<td>’</td>
<td>7</td>
<td>G</td>
<td>W</td>
<td>g</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>BS</td>
<td>CAN</td>
<td>(</td>
<td>8</td>
<td>H</td>
<td>X</td>
<td>h</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>HT</td>
<td>EM</td>
<td>)</td>
<td>9</td>
<td>I</td>
<td>Y</td>
<td>i</td>
<td>y</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>LF</td>
<td>SUB</td>
<td>:</td>
<td>J</td>
<td>Z</td>
<td></td>
<td>j</td>
<td>z</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>VT</td>
<td>EBD</td>
<td>;</td>
<td>K</td>
<td></td>
<td></td>
<td>k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>FF</td>
<td>FS</td>
<td>&lt;</td>
<td>L</td>
<td>\</td>
<td></td>
<td>l</td>
<td>\</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>CR</td>
<td>GS</td>
<td>=</td>
<td>M</td>
<td>]</td>
<td>m</td>
<td>]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>SD</td>
<td>RS</td>
<td>&gt;</td>
<td>N</td>
<td>^</td>
<td>n</td>
<td>^</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>SI</td>
<td>US</td>
<td>?</td>
<td>O</td>
<td>_</td>
<td>o</td>
<td>DEL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Robert Kelly, 2001-2018
Modern Approach to Encoding

- Establish
  - Universal set of characters that can be encoded in a variety of ways
  - Ordering of the characters
- Character repertoire - the full set of abstract characters that a system supports, and might allow
  - No additions - e.g., ASCII
  - Additions
- Examples
  - Unicode
  - ISO/IEC10646

Unicode

- Can represent the characters of every language in the world
- Contains
  - more than 110,000 characters (Universal Character Set)
  - 100 scripts (e.g., Latin, Arabic) These code points are the HTML numeric references
  - Codepoint for every character
  - A 6-part codespace (e.g., Western alphabet codes)
- Equivalent (almost) to ISO 10646
- Implemented by various encodings
  - UTF-8 - one byte for ASCII characters and up to 4 bytes for other characters
  - UTF-16 - 2-4 bytes for each character
- Java uses Unicode as its default character set
Unicode Codespace Allocation

- The lowest-numbered Unicode characters comprise the ASCII code - preserves backwards compatibility

<table>
<thead>
<tr>
<th>Character Types</th>
<th>Language</th>
<th>Number of Characters</th>
<th>Hexadecimal Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabets</td>
<td>Latin, Greek, Cyrillic, etc.</td>
<td>8192</td>
<td>0000 to 1FFF</td>
</tr>
<tr>
<td>Symbols</td>
<td>Dingbats, Mathematical, etc.</td>
<td>4096</td>
<td>2000 to 2FFF</td>
</tr>
<tr>
<td>CJK</td>
<td>Chinese, Japanese, and Korean phonetic symbols and punctuation.</td>
<td>4096</td>
<td>3000 to 3FFF</td>
</tr>
<tr>
<td>Han</td>
<td>Unified Chinese, Japanese, and Korean</td>
<td>40,960</td>
<td>4000 to DFFF</td>
</tr>
<tr>
<td>Han Expansion</td>
<td></td>
<td>4096</td>
<td>E000 to EFFF</td>
</tr>
<tr>
<td>User Defined</td>
<td></td>
<td>4095</td>
<td>F000 to FFFE</td>
</tr>
</tbody>
</table>

Example - HTML

- An HTML document consists of Unicode characters
- When transmitted, the document is encoded according to document / server instructions, as in `<meta charset=\"UTF-8\" />`
- When the encoding or editor does not support all the Unicode characters used in the document, characters can be escaped using an entity reference

<table>
<thead>
<tr>
<th>Entity Reference</th>
<th>Category</th>
<th>Displays As</th>
</tr>
</thead>
<tbody>
<tr>
<td>&amp;#x5E7;</td>
<td>Hebrew</td>
<td>ם</td>
</tr>
<tr>
<td>&amp;#x645;</td>
<td>Arabic</td>
<td>﷢</td>
</tr>
<tr>
<td>&amp;#x8449;</td>
<td>Chinese</td>
<td>葉</td>
</tr>
<tr>
<td>&amp;#x5B5B;</td>
<td>Korean</td>
<td></td>
</tr>
</tbody>
</table>
Special Characters

Characters can be used directly or as a special reference (if it is not in the character set or if there is a “meaning conflict”

Character references can be numeric or literal

You should replace
Copyright© 1996-2004
with
Copyright© 1996-2004

Literal character reference

Good practice to add the symbol when it is missing

Character References

Numeric references (decimal or hexadecimal)

- &lt;29; - å (Norwegian)
- &lt;x6C34; - 水 (Chinese character for water)

Character entity references

- &gt; represents the > sign

Ariel Unicode MS font supports Unicode characters

Numeric references refer to Unicode, which is then mapped into the specific encoding (e.g., UTF-8)

Unicode is like a virtual encoding
Body Content View Descriptions

- An HTML page can describe some of the styling information in external style sheets
  ```
  <link rel="stylesheet" href="original_files/nav.css" type="text/css" />
  <link rel="stylesheet" href="original_files/right.css" type="text/css" />
  <link rel="stylesheet" href="original_files/calander.css" type="text/css" />
  ```

- Style information is usually applied to the element (e.g., td) or to enclosed elements (e.g., with font). More on this in the next class session

Additional HTML Data Types

- Colors
  - attribute value type "color" refers to color sRGB definitions
  - A color value may either be a hexadecimal number (prefixed by a hash mark) or one of sixteen color names

- Length - pixels or percentage

- Media descriptors
  - Screen, tty, tv, projection, print, handheld, print, Braille, aural, all
Tables

- The HTML table model allows authors to arrange data (text, preformatted text, images, links, forms, form fields, other tables, etc.) into rows and columns of cells
- Tables should resize dynamically
- Should allow incremental display
- Allow head, foot, and body groupings
- Cells can span multiple rows and columns

Most html pages use tables to organize the content on the page (including embedded tables)

Forms

- A form element usually contains text, along with GUI components and a submit button
- Typical GUI components
  - Text box (input element, with type of text)
  - Dropdown (select element)
  - Check box (input element, with type of checkbox)
  - Radio button (input element, with type of radio)
Form Example

```html
<form method="post" action="Mets/tix" >
  <input name="Team" value="New York Mets" type="hidden" />
  ...
  <div align="right">Opponent:</div>
  <input name="Opponent" size="20" class="nav" type="text" />
  <div align="right">Date:</div>
  <input name="Date" size="10" class="nav" type="text" />
  mm/dd/yy
  <div align="right">*Number of tickets:</div>
  <select name="Number" class="nav">
    <option selected="selected">Select</option>
    <option>1</option>
    <option>2</option>
    <option>3</option>
    <option>4</option>
    <option>5+</option>
  </select>
  ...
</form>
```

HTML5 Features

- Both xml and html syntax included
- New features (e.g., video and audio)
- Enriched semantic structure (e.g., header, section, and article)
- Reduced dependence on div and span
- Well-defined handling of incorrect syntax
- New and extended APIs (e.g., DOM)

DOM API is now a part of the HTML spec
New HTML5 Structure Elements

- `<section>` - sections of pages
- `<header>` - header of a page
- `<footer>` - footer of a page
- `<nav>` - navigation on a page
- `<article>` - article or primary content on a page
- `<aside>` - extra content like a sidebar on a page
- `<figure>` - images that annotate an article

Are you familiar with these?

Is this an aside?

---

HTML5 Features

- New form elements - datetime, datetime-local, date, month, week, time, number, range, email, url
- New elements
  - `<canvas>` - gives you a drawing space in JavaScript on your Web pages
  - `<video>` - add video
  - `<audio>` - add sound
- Removes elements - many of them replaced by CSS and already deprecated
Session 3 - HTML

HTML5 APIs

- Improved APIs can be used with JavaScript
  - DOM - Document Object Model
- New
  - Dynamic rendering of 2D shapes and bitmap images
  - Cross document messaging
  - Microdata - embeds metadata within page content
- Separate from HTML5
  - Web storage - similar to cookies, but with enhanced capacity
  - Geolocation

Have You Satisfied the Lecture Objectives

- Become familiar with HTML syntax
- Understand the relationship between an HTML document and the corresponding element tree
- Understand the evolution of HTML