Session 3

HTML

Tim Berners-Lee

1

Reading

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

  en.wikipedia.org/wiki/Character_encodings_in_HTML

2

References

- WWW Consortium - HTML 5

  https://www.w3.org/TR/html5/
- HTML character entity references

  www.htmlhelp.com/reference/html40/entities/

3

Lecture Objectives

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

4

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 of the evolution of many Internet standards and technologies

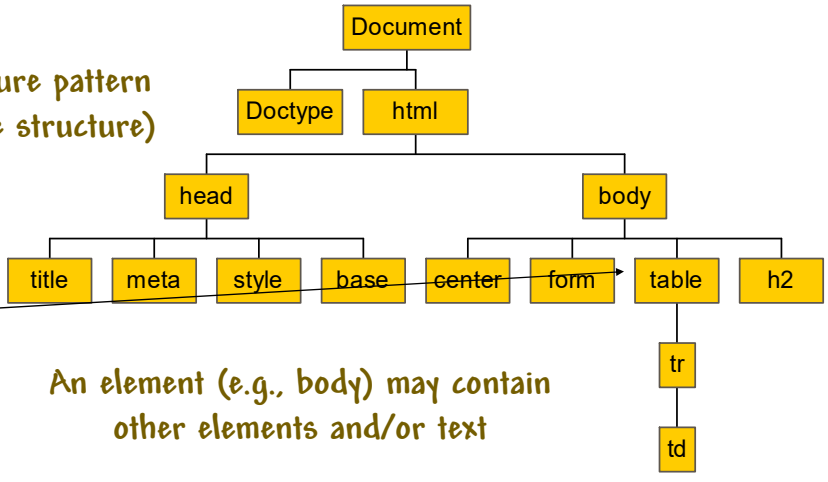
5

Example Document Structure

- An HTML page is a tree of html elements

Note the text enclosure pattern (that defines the tree structure)

```
...
<body>
  <table>
    ...
  </table>
...
</body>
```



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

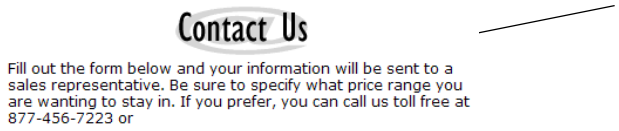
6

HTML Element

- An element consists of a begin tag and an end tag

Notice the distinction between an element and a tag

```
<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>
```



This html produces this browser display

7

What are the Components of HTML?

tag

```
<html>
```

element name

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

- Must be properly nested
- An end tag closes all intervening tags

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

```
<body> and </BODY>
```

Element names are case insensitive in html

8

What are the Components of HTML?

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

9

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

10

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

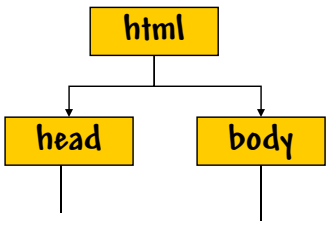
Doctype

```
<!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




```
<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>
```

MIME type

Style sheet

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)

15

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)

16

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>
```

This is a paragraph tag

- This is an item in an unordered list.
- This is another item in that list.

1. This is an item in an ordered list.
2. This is another item in that list.

Definition lists are also available in html

Be sure that you understand the html tags for ordered and unordered lists

17

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)

18

Characters

```
<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

19

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"

20

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)

21

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

22

Character Code Issues

- Character codes
  - Mapping of characters to strings of binary digits
  - E.g., "S" usually is usually mapped to 01000011$_2$
- 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"

23

ASCII Reference Table

Note the ordering of characters

MSD / LSD	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P		p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACJ	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

74_{16}
1110100_2

24

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

25

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)
  - Codepoint for every character
  - A 6-part codespace (e.g., Western alphabet codes)

These code points are the HTML numeric references

- 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

26

Unicode Codespace Allocation

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

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

27

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" />
```

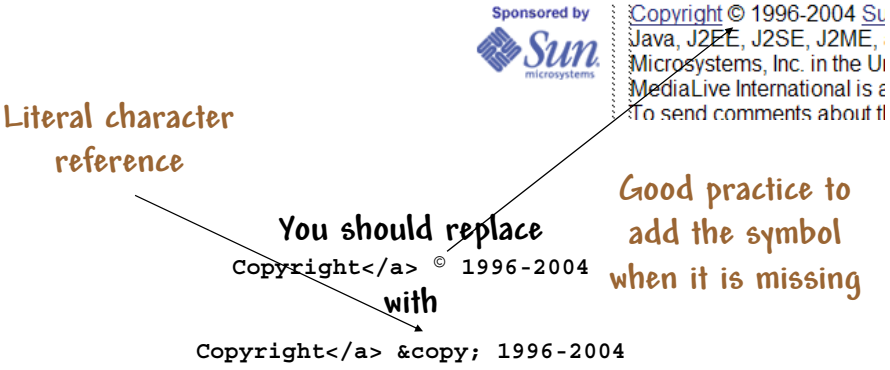
- When the encoding or editor does not support all the Unicode characters used in the document, characters can be escaped using an entity reference

Entity Reference	Category	Displays As
ק	Hebrew	ק
م	Arabic	م
葉	Chinese	葉
떫	Korean	떫

28

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



29

Character References

- Numeric references (decimal or hexadecimal)
  - å - å (Norwegian)
  - 水 - 水 (Chinese character for water)
- Character entity references
  - > 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)

Numeric references use either decimal notation (#nnnn) or hex notation (#xhhhh), with or without leading zeroes

Unicode is like a virtual encoding

30

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

31

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

32

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)

It is usually not a good idea to use exact table (e.g., column) pixel dimensions

33

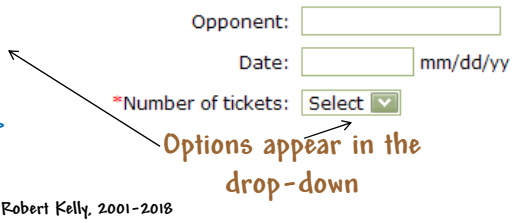
Forms

- 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)

34

Form Example

```
<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>
```



Options appear in the drop-down

35

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

36

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?

37

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

38

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

39

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

40