HTML, XML, and Web Search
Needles in the Haystack: Google and other brokers in the bits bazaar
Everything we know is online

- Newspapers
- Scientific articles
- Books
- Encyclopedias (Wikipedia)
- Dictionaries and Thesauruses
- Art
Why the WWW is so brilliant

• WWW = the servers + the data
• All computers use one standard protocol (HTTP)
  – Everyone can communicate
• Publishing and accessing information is decentralized
  – No limits on what is possible to say or search for
• Critical mass of data
Seeing Other People's Data

WWW exploits one protocol, neutralizes differences at endpoints.
Bits and bytes encode information, but that's not all

**Tags** encode format and some structure in word processor documents

**Tags** encode format and some structure in HTML

**Tags** are one form of metadata

**Metadata** is information about information
Shiba Inus Meme

"Doge" is a slang term for "dog" primarily associated with pictures of the Shiba Inus breed.

What a beautiful doge!
A Webpage - Metadata

- Title
- Description
- Font names
- Structure

```html
<html lang="en">
<head>
  <link href='http://fonts.googleapis.com/css?family=Gabriela' rel='stylesheet' type='text/css'>
  <title>Doge</title>
  <meta name="description" content="Doge meme page">
</head>
<body>
<h1>Shiba Inus Meme</h1>
<p>"Doge" is a slang term for "dog" primarily associated with pictures of the Shiba Inus breed.</p>
<p><img src="http://i.imgur.com/sUtamDC.jpg" width="200" alt="SUCH MEME" /></p>
<p>What a beautiful doge!</p>
</body>
</html>
```
HTML and the Web

• Web uses **http://** protocol

• Browser asks for a web page
  
  – Usually expressed in *hyper-text markup language*, HTML
    
    • Hyper-text means you can have links that leave linear stream of text, let user leave and come back
    
    • Markup language is a notation to describe how a published document looks
      
      – fonts
      
      – text color
      
      – heading
      
      – images
      
      – page layout
Basics of HTML #1

• Rule 1
  – Content is given directly; anything not content is given inside tags

• Rule 2
  – Tags are made of < and > used like this:
    <p style="color:red">This is a paragraph</p>
    Start tag                  Content                  End tag

• Rule 3
  – Tags must be paired (start-end) or "self-terminated"
Introduction to HTML

When you look at a web page in a web browser, you see, at the simplest level, words. These words usually have some style of different fonts, font sizes and colors. In many cases a page also displays images or maybe video. Sometimes there is a form search) for information, or customize the display of the page to your liking. Often a page contains content that moves or changes, the page remains the same.

Several technologies (such as CSS, JavaScript, Flash, AJAX, JSON) can be used to define the elements of a web page. However, the page is defined using HTML (HyperText Markup Language). Without HTML, there is no web page. HTML is the shell that holds international standard whose specifications are maintained by the World Wide Web Consortium (W3C) and the Web Hypertext Working Group (WHATWG). The WHATWG considers HTML a "living standard" which is constantly evolving, whereas the W3C HTML, the most current of which is HTML5 and on the evolution of HTML (HTML 5.1).
Live Coding

- http://codepen.io/
• Rule 4
  – An HTML file has this structure

```html
<html>
  <head><title>Name of Page</title></head>
  <body>
    Actual contents here
  </body>
</html>
```

• Rule 5
  – Tags must be properly nested
● Rule 6
  – Whiteline is mostly ignored

● Rule 7
  – Attributes \(\text{style} = \text{color:red}\)) are preceded by space, name is not quoted, value is quoted
Example

- Write HTML using a text editor
  - Notepad++
  - TextWrangler
  - Geany
- File extension is .html
- Open using Chrome or Firefox to see it

Mary Had a Little Lamb by Vi Hart (Lyrics to 12-tone arrangement for vocal quartet)

Mary had a little lamb
little lamb
little lamb
Mary had a little lamb
it's fleece was white as snow.

And everywhere that Mary went, Mary went, Mary went, everywhere that Mary went,
Yes everywhere that Mary went her lamb was sure to go.

Then someone burned down Mary’s house, someone burned down Mary’s house,
She'd been dead a hundred years, hundred years, hundred years, she'd been dead for more.
<!DOCTYPE html>
<html>
<head><title>Mary Had a Little Lamb (Vi Hart)</title></head>
<body>
<img src="http://i.imgur.com/Wf5qms5.png" />
<p>Mary Had a Little Lamb by Vi Hart (Lyrics to 12-tone arrangement)

Mary had a little lamb,
<br/>little lamb
<br/>little lamb
<br/>Mary had a little lamb
<br/>it's fleece was white as snow.

And everywhere that Mary went,
Mary went, Mary went,
everywhere that Mary went
the lamb was sure to go.
Yes everywhere that Mary went her lamb was sure to go.

Then someone burned down Mary's house, someone burned down the house
She'd been dead a hundred years,
hundred years,
hundred years,
she'd been dead a hundred years,
her lambs were dead for more.

</p>
</body>
</html>

http://codepen.io/anon/pen/wnBog
• To put in an image:
  
  `<img src="tones.png" alt="Twelve Tones" />`

• To put in a link:
  
  `<a href="http://google.com/">Google</a>`

• Lots of good tutorials online
  
  - http://www.w3schools.com/html/default.asp
Image Servers

• You could also upload your image to an *image server*
  – e.g. *imgur.com*
  – Then link directly to image URL on image server

• What are you giving up when you do this?
  – What are the tradeoffs?
Homework 8: Making a Personal Webpage

To Turn In. Submit the link to your html webpage under “Homework 8” in eCommons.

Step 1. Gaining access to your user account

If you are working on your personal machine, in Step 4, you will need to download FileZilla and follow the instructions (at the end of the homework) for ‘file transfer protocol’, or simply ftp. This is a method of remotely accessing any files saved to your user account on the school servers.

Alternatively, if you are working on the lab machines, you do NOT need to use FileZilla since you will be able to modify and save files directly to your account.

Step 2. Modifying HTML

Modify the HTML source code from the website http://people.rose-hulman.edu/~plherron/ according to
What are you supposed to learn?

- HTML lets you programmatically indicate how particular content should be displayed.
- It can be served up by any HTTP server anywhere in the world.
- Typically uses UTF-8 encoding to guarantee compatibility.
End