LAST WEEK ON IO LAB

Project 2 demonstrations on controlled vocabularies.

Semantic web technologies including microformats and RDF.

Introduction to Google App Engine.
THE SEMANTIC WEB
Information made for machines
RDF @ MICROSOFT

Thanks, Tanushree!
RDFa
RDF in attributes

<p xmlns:dc="http://purl.org/dc/elements/1.1/"
    about="http://www.example.com/books/wikinomics">
    In his latest book
    <cite property="dc:title">Wikinomics</cite>,
    <span property="dc:creator">Don Tapscott</span>
    explains deep changes in technology,
    demographics and business.
    The book is due to be published in
    <span property="dc:date" content="2006-10-01">October 2006</span>.
</p>
SCREEN SCRAPING
http://microformats.org
<div class="vcard">
  <a class="url fn" href="http://ryangreenberg.com">
    Ryan Greenberg
  </a>
  <a class="email" href="mailto:ryan@ischool.berkeley.edu">
    ryan@ischool.berkeley.edu
  </a>
  <div class="adr">
    <div class="street-address">2020 Oregon St</div>
    <span class="locality">Berkeley</span>,
    <span class="region">CA</span>
    <span class="postal-code">94703</span>
  </div>
  <div class="tel">510-843-6929</div>
</div>
XFN
XHTML Friends Network

<a rel="friend co-worker met" href="http://npdoty.name">Nick</a>
<a rel="sweetheart friend met" href="http://www.facebook.com/profile.php?id=643667604">Caitlin</a>
<a rel="crush" href="http://www.beyonceonline.com">Beyoncé</a>
<a rel="enemy" href="http://e.ntangled.net">Michael</a>
XFN
XHTML Friends Network

rel="me"
WIKIPEDIA MICROFORMATS

GEO MICROFORMAT

```html
<div class="geo">
  <span class="latitude">37.87167</span>,
  <span class="longitude">-122.27278</span>
</div>

<div class="geo">
  <abbr class="latitude" title="37.87167">37°52'18"</abbr>,
  <abbr class="longitude" title="-122.27278">122°16'22"</abbr>
</div>
```
GOOGLE RICH SNIPPETS
HTML PARSING

- BeautifulSoup (and soupselect) in Python on App Engine
- lxml in Python
- Nokogiri in Ruby
- jQuery in JavaScript
RegEx match open tags except XHTML self-contained tags

You can't parse [X]HTML with regex. Because HTML can't be parsed by regex. Regex is not a tool that can be used to correctly parse HTML. As I have answered in HTML-and-regex questions here so many times before, the use of regex will not allow you to consume HTML. Regular expressions are a tool that is insufficiently sophisticated to understand the constructs employed by HTML. HTML is not a regular language and hence cannot be parsed by regular expressions. Regex queries are not equipped to break down HTML into its meaningful parts. so many times but it is not getting to me. Even enhanced irregular regular expressions as used by Perl are not up to the task of parsing HTML. You will never make me crack. HTML is a language of sufficient complexity that it cannot be parsed by regular expressions. Even Jon Skeet cannot parse HTML using regular expressions. Every time you attempt to parse HTML with regular expressions, the unholy child weeps the blood of virgins, and Russian hackers pwn your weapp. Parsing HTML with regex summons taint souls into the realm of the living. HTML and regex go together like love, marriage, and ritual infanticide. The &lt;center&gt; cannot hold it is too late. The force of regex and HTML together in the same conceptual space will destroy your mind like so much wetty putty. If you parse HTML with regex you are giving in to Them and their blasphemous ways which doom us all to inhuman toil for the One whose Name cannot be expressed in the Basic Multilingual Plane, he comes. HTML-plus-regexp will liquify the nerves of the sentient whilst you observe, your psyche withering in the onslaught of horror. RegEx"x-based HTML parsers are the cancer that is killing StackOverflow it is too late it is too late we cannot be saved the trangression of a child ensures regex will consume all living tissue (except for HTML which it cannot, as previously prophesied) dear lord help us how can anyone survive this scourge using regex to parse HTML has doomed humanity to an eternity of dread torture and security holes using regex as a tool to process HTML establishes a breach between this world and the dread realm of corp rup entities (like SGML entities, but more corrupt) a mere glimpse of the world of regex parsers for HTML will instantly transport a programmer's consciousness into a world of ceaseless screaming, he comes, the pestilent-slimy regex-infection will devour your HTML parser, application and existence for all time like Visual Basic only worse he comes he comes do not fight he comes his inhuman radiance destroying all enlightenment, HTML tags leaking from your eyes like liquid pain, the song of regular expression parsing will extinguish the voices of mortal man from the sphere I can see it can you see it is beautiful the final snuffing of the lies of Man ALL IS LOS...
PARSING HREVIEV
APP ENGINE REDUX
OVERVIEW OF PROJECT 3 IDEAS

http://courses.ischool.berkeley.edu/i290-4/f09/projects.php#assignment-3
TRIPLE STORE + QUERIES
CREATE OR CONSUME
ISCHOOL + FOAF
ISCHOOL + MICROFORMATS
FREEBASE + RDF
MICROFORMATS
FOR NEXT WEEK

Decide on an idea for project 3, coordinating with other groups if necessary.

Student presentations on CSS tips and tricks in Thursday’s Lab Lab.