

# Using Metadata to Improve Search Results

Jeremy Ey  
March 12th, 2007  
CSC 6800-001

# Background Information

- Hard to create machine understand of html.
- Use of common structures to provide information about what is contained on the page.
- Microformats provide a common set of html class names to denote this information

# Client Side Use Examples

- Example of an hCard within a webpage:

<http://michael-mccracken.net/wp/>

```
1. <div class="vcard" style="padding: 15px; font-color: 0xd9c19f;">
2.   
3.
4.   <span class="fn n"> <span class="given-name">Michael</span>
5.
6.     <span class="additional-name">O</span>
7.     <span class="family-name">McCracken</span>
8. </span>
9.
10. <div class="org">UCSD CSE</div>
11. <a class="email" href="mailto:...">email</a>
12. <div class="adr">
13.   <span class="locality">San Diego</span>, <span class="region">CA</span>
<span class="postal-code">92122</span>
14.
15. </div>
16. <a class="url" href="aim:goim?screenname=q606">AIM</a><br/>
17. <a class="url" href="http://michael-mccracken.net/">Home page</a><br/>
18. <a class="url" href="http://michael-mccracken.net/wp/">Weblog</a><br/>
19. <a class="url" href="http://www.cs.ucsd.edu/~mmccrack/">School page</a>
20. </div>
```

# GRDDL

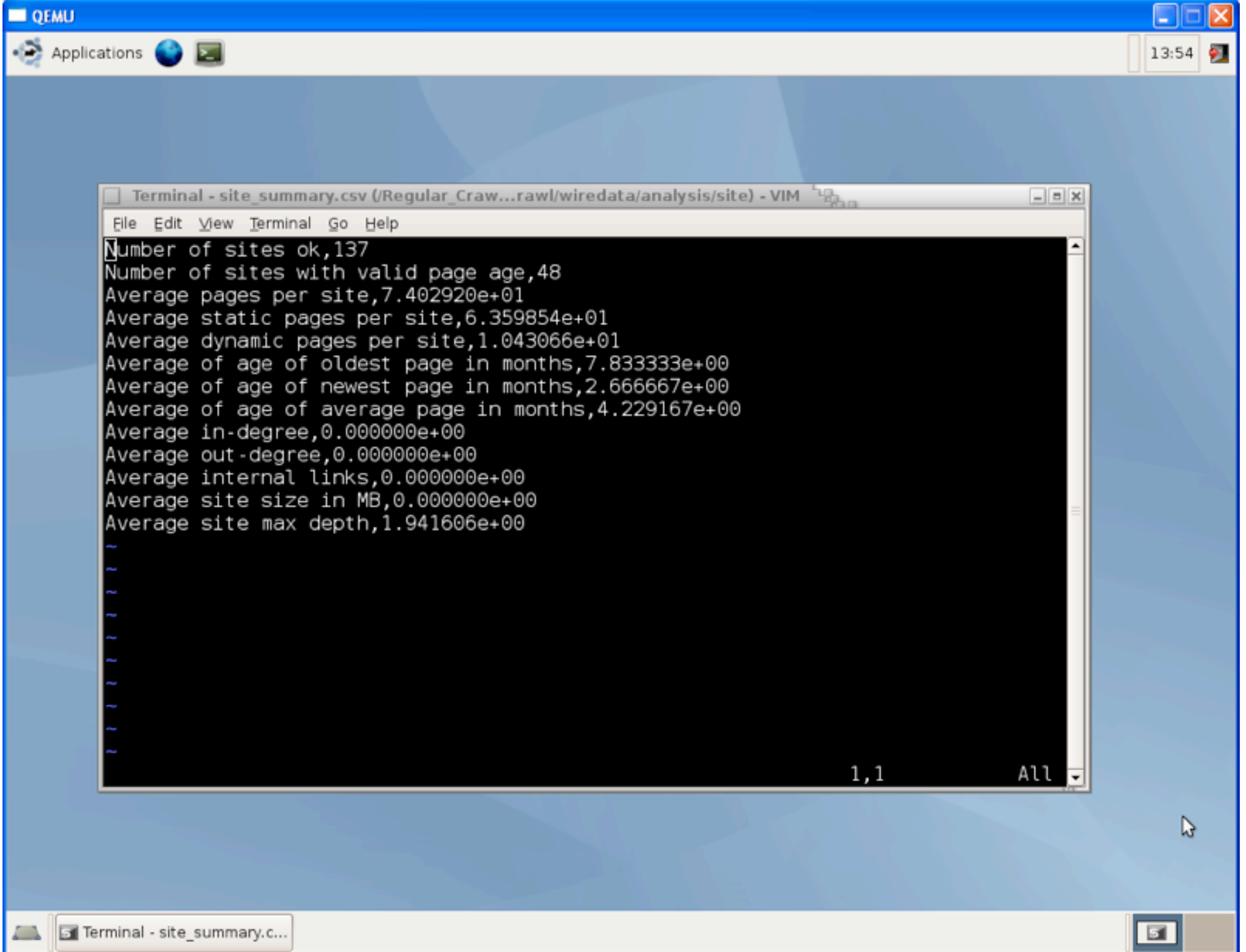
- Gleaning Resource Descriptions from Dialects of Languages
- Provide references to xslt stylesheets to transform content from one format to another.

# Setup

- QEMU Virutal Machine with xubuntu
- run on Windows XP hosts
- WIRE
- Local DNS Cache

# WIRE Configuration

- looking at .com .net .edu .org and .us domains
- maximum of 10K urls per site
- maximum depth of 5 levels
- maximum of 50K documents per harvester run



# Starting URLs

- <http://www.csc.tnitech.edu/>
- <http://microformats.org/wiki/hcard-examples-in-wild/>
- <http://michael-mccracken.net/wp/>

# Crawling Results

- 6 rounds of harvesting
- 800 sites
- 13436 urls

# Additional Results

- I'm starting to question the determinism of computers

# Next Steps

- Count and index hCard data
- Interface to access hCard Index
- Expand metadata types reviewed  
(GRDDL)

# Future Work

- Adjust crawling depth to further explore sites where data has previously been found
- Support plugins from other systems to expand data collection ability