

Web Scraping: Room Full of Interns in a Box



What it's good for

- For a Kansas City Star series on the insurance industry, I scraped 35 million insurance complaint data from the National Association of Insurance Commissioners.
- At the AJC, a scrape of the state legislature's Web site resulted in a story about a legislator who missed more than 90 percent of the votes ... because he was caught in an FBI money laundering sting and had agreed not to cast votes while acting as an informant in a corruption investigation.
- The same scrape provided legislative reporters a daily spreadsheet to track bills.
- Arlen Poort, NRC Handelsblad, is scraping a site with data about home currently for sale in The Netherlands to track length of time on market, asking prices, sale prices and other indicators that might signal trend in the market.

What it's good for

- Instead of paying \$1,000 a year for county assessor data, Agustín Armendariz scrapes the assessor's web site daily to capture house sales and feed them directly to a search/mapping application on the San Diego Union-Tribune Web site.
- Ian Dempsy, the Tacoma, WA, News-Tribune, scrapes the local sheriff's Web site hourly to collect jail bookings.
- Sarah Cohen, Washington Post, avoided months of FOIA delays by scraping corn subsidy data from the USDA site for a series on ag subsidies.
- For the same series, data scraped from another USDA site helped Sarah find insurance agents who had received disaster payments.
- Sarah, again, scraped the US Trade Commission site to get data for a story on special tariff exemptions passed into law as earmarks.

Jennifer LaFluer, Dallas Morning News, scraped soil sample data from the EPA Web site for a story about soil contamination after Hurricane Katrina

KATRINA'S TOXIC LEGACY

The New Orleans flood that followed Hurricane Katrina left toxic substances when it receded. Metals, industrial compounds, petroleum products and other materials coated many neighborhoods, tests by the Environmental Protection Agency show.

The *Dallas Morning News* compared the amounts the EPA found with the screening levels that the agency uses to look for potential toxic problems in residential soil. These maps show where tests found contamination levels above those screening levels, as measured in parts per million.

These results alone don't prove that any particular neighborhood faces a bigger or smaller risk than other neighborhoods. More tests and an extensive scientific study would be needed to show that.

However, they do show where the tests revealed the highest levels. Bigger symbols represent greater contamination, not larger areas of contamination.

ARSENIC: A toxic and cancer-causing metal.



BENZO(A)PYRENE, DIBENZO(A,H)ANTHRACENE: Two chemicals known as polycyclic aromatic hydrocarbons. They are carcinogens.

Benzo(a)pyrene
0.09 - 1.68 1.69 - 4.24 4.25 - 17.70 17.71 - 35.50

Dibenzo(a,h)anthracene
0.11 - 0.22 0.23 - 0.72 0.73 - 2.58 2.59 - 6.43



DIELDRIN: A cancer-causing pesticide used against termites, banned since 1987.



TOTAL PETROLEUM HYDROCARBONS AND DIESEL RANGE ORGANICS: Components of oil, gasoline, diesel fuel.

Total petroleum hydrocarbons

138 - 5,270 5,270 - 14,900 14,900 - 33,600 33,600 - 1,600,000

Diesel range organics

400 - 711 711 - 1,890 1,890 - 5,360 5,360.01 - 15,600



LEAD: A powerful poison that can seriously damage children's development.



SOURCES: Environmental Protection Agency data; *Dallas Morning News* research.

HTTP: Browsers Talking With Servers

Browser => HTTP request => Server

<http://bolles.ire.org/dij/by-state.ptmpl?sn=FL>

GET /dij/by-state.ptmpl?sn=FL HTTP/1.1

Host: bolles.ire.org

User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.14) Gecko/20080404

Firefox/2.0.0.14

Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5

Accept-Language: en-us,en;q=0.5

Accept-Encoding: gzip,deflate

Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7

Keep-Alive: 300

Connection: keep-alive

Referer: <http://bolles.ire.org/dij/usa.html>

HTTP: Browsers Talking With Servers

Browser <= HTTP response <= Server

HTTP/1.x 200 OK

Date: Thu, 05 Jun 2008 04:21:43 GMT

Server: Apache/1.3.27 (Unix) mod_fastcgi/2.4.0

Expires: Thu, 05 Jun 2008 04:26:43 GMT

Content-Type: text/html; charset=ISO-8859-1

X-Cache: MISS from localhost

Connection: close

HTML: The Document Body

```
<b>Name:</b></td><td>Joe Adams</td></tr>
<tr><td align="right"><b>Affiliation:</b></td><td><a href="by-
affiliation.ptmpl?affiliation=The%20Florida%20Times-Union">The Florida Times-Union</a></td></tr>
<tr><td align="right"><b>Address:</b></td><td>Jacksonville <a href="by-
state.ptmpl?sn=FL">FL</a> 32202 <a href="by-country.ptmpl?country=USA">USA</a></td></tr>

<tr><td align="right"><b>Phone:</b></td><td>(904) 3594534</td></tr>
<tr><td align="right"><b>Interests:</b></td><td><a href="by-
interest.ptmpl?interest=First%20Amendment%2FFOIA">First Amendment/FOIA</a>, <a href="by-
interest.ptmpl?interest=Local%20Government%2FCity%20Hall">Local Government/City Hall</a>, <a
href="by-interest.ptmpl?interest=State%20Government">State
Government</a></td></tr></tbody></table></p><p>
<table border="0" cellpadding="2">
<tbody><tr><td align="right"><b>Name:</b></td><td>Steve Andrews</td></tr>
<tr><td align="right"><b>Affiliation:</b></td><td><a href="by-affiliation.ptmpl?affiliation=NBC%20-
%20WFLA-TV">NBC - WFLA-TV</a></td></tr>

<tr><td align="right"><b>Address:</b></td><td>Tampa <a href="by-state.ptmpl?sn=FL">FL</a>
33511 <a href="by-country.ptmpl?country=USA">USA</a></td></tr>
<tr><td align="right"><b>Phone:</b></td><td>(813) 221-5779</td></tr>
```

Complicating factors

- JavaScript: Code is executed by the browser on the client side. Requires a JavaScript engine to interpret the code.
- AJAX: Asynchronous requests and responses
- HTTPS: Encryption
- Denial of service attack protection
- Header User-Agent and Referrer checkers
- CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart
- Data validation
- Scrapers break

Tools

- The analyze a Web site: Firefox
 - Live HTTP Header
 - Web Developer
 - Firebug
- Scripting languages: Perl Python, Ruby
- Spidering helpers:
 - The old way: LWP
 - The new way: Mechanize
- Parsing helpers
 - The old way: regular expressions: `/<td>(.)</td>`
 - The new way: `Table::Extract`, `TokenParser`, `TreeBuilder`, `XML::DOM`

Precautions

- If it's a private site, be respectful, if not blindly obedient to its terms of use agreement and robot.txt.
- Public sites are open game.
- Public or private, respect the server's bandwidth
 - Rule of thumb: no more than 2 requests/second
- Cache or download parts of the site for testing