ScraperWiki Tutorial

Levine, Thomas
2012

Dostupný z http://www.nusl.cz/ntk/nusl-127015

Tom brags about his data-cleaning projects, and you can too.

Outline

Bragging about projects

- New Orleans Wetlands
- Financial Products in Africa
- Sampling frame for a study on toilet-use
- Middle names

Thoughts

New Orleans Wetlands

New Orleans Wetlands

The Gulf Restoration Network on wetlands

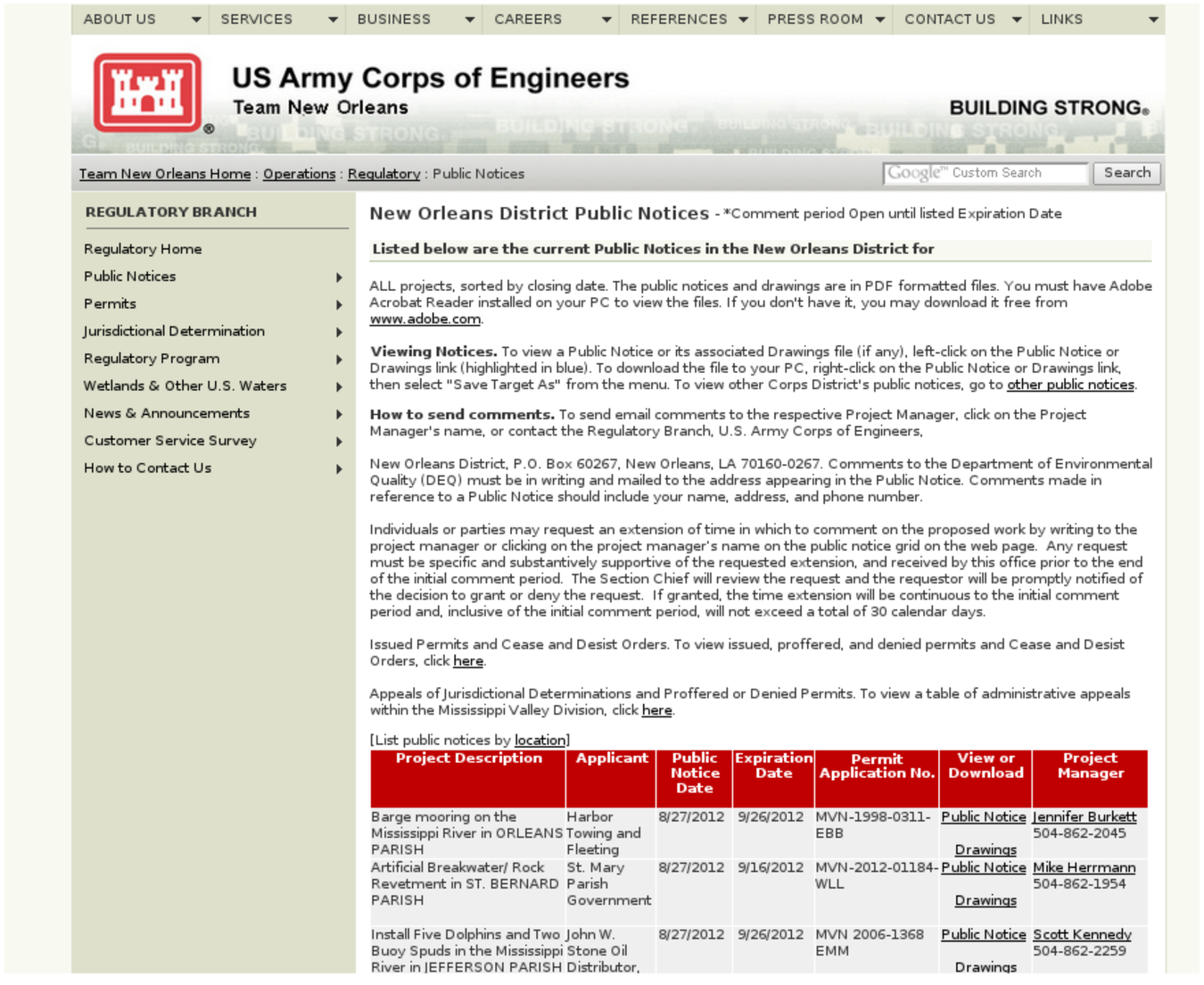
> *The [Gulf Restoration Network] works to protect wetlands from reckless development, destructive logging practices, and harmful U.S. Army Corps of Engineers projects and policies.*

More specifically

- You apply for a permit build on the wetlands in the United States.
- The Army Corps of Engineers can approve permits that meet certain criteria.
- I have no idea how the Army got this job.
- The Army, understandably, doesn't scrutinize permits as much as the Gulf Restoration Network would like.

New Orleans Wetlands

Applications get posted to a website.



ABOUT US | SERVICES | BUSINESS | CAREERS | REFERENCES | PRESS ROOM | CONTACT US | LINKS

US Army Corps of Engineers
Team New Orleans
BUILDING STRONG®

Team New Orleans Home : Operations : Regulatory : Public Notices

Google™ Custom Search | Search

REGULATORY BRANCH

- Regulatory Home
- Public Notices
- Permits
- Jurisdictional Determination
- Regulatory Program
- Wetlands & Other U.S. Waters
- News & Announcements
- Customer Service Survey
- How to Contact Us

New Orleans District Public Notices - *Comment period Open until listed Expiration Date

Listed below are the current Public Notices in the New Orleans District for

ALL projects, sorted by closing date. The public notices and drawings are in PDF formatted files. You must have Adobe Acrobat Reader installed on your PC to view the files. If you don't have it, you may download it free from www.adobe.com.

Viewing Notices. To view a Public Notice or its associated Drawings file (if any), left-click on the Public Notice or Drawings link (highlighted in blue). To download the file to your PC, right-click on the Public Notice or Drawings link, then select "Save Target As" from the menu. To view other Corps District's public notices, go to other public notices.

How to send comments. To send email comments to the respective Project Manager, click on the Project Manager's name, or contact the Regulatory Branch, U.S. Army Corps of Engineers,

New Orleans District, P.O. Box 60267, New Orleans, LA 70160-0267. Comments to the Department of Environmental Quality (DEQ) must be in writing and mailed to the address appearing in the Public Notice. Comments made in reference to a Public Notice should include your name, address, and phone number.

Individuals or parties may request an extension of time in which to comment on the proposed work by writing to the project manager or clicking on the project manager's name on the public notice grid on the web page. Any request must be specific and substantively supportive of the requested extension, and received by this office prior to the end of the initial comment period. The Section Chief will review the request and the requestor will be promptly notified of the decision to grant or deny the request. If granted, the time extension will be continuous to the initial comment period and, inclusive of the initial comment period, will not exceed a total of 30 calendar days.

Issued Permits and Cease and Desist Orders. To view issued, proffered, and denied permits and Cease and Desist Orders, click here.

Appeals of Jurisdictional Determinations and Proffered or Denied Permits. To view a table of administrative appeals within the Mississippi Valley Division, click here.

[List public notices by location]

Project Description	Applicant	Public Notice Date	Expiration Date	Permit Application No.	View or Download	Project Manager
Barge mooring on the Mississippi River in ORLEANS PARISH	Harbor Towing and Fleeting	8/27/2012	9/26/2012	MVN-1998-0311-EBB	Public Notice Drawings	Jennifer Burkett 504-862-2045
Artificial Breakwater/ Rock Revetment in ST. BERNARD PARISH	St. Mary Parish Government	8/27/2012	9/16/2012	MVN-2012-01184-WLL	Public Notice Drawings	Mike Herrmann 504-862-1954
Install Five Dolphins and Two Buoy Spuds in the Mississippi River in JEFFERSON PARISH	John W. Stone Oil Distributor,	8/27/2012	9/26/2012	MVN 2006-1368 EMM	Public Notice Drawings	Scott Kennedy 504-862-2259

New Orleans Wetlands

Applications look like this.

DEPARTMENT OF THE ARMY
NEW ORLEANS DISTRICT, CORPS OF ENGINEERS
P.O. BOX 60267
NEW ORLEANS, LOUISIANA 70160-0267

REPLY TO
ATTENTION OF

August 27, 2012

Operations Division
Eastern Evaluation Section

(504) 862-2045
Project Manager
Jennifer Burkett
SUBJECT: MVN 1998-0311 EBB

PUBLIC NOTICE

Interested parties are hereby notified that a permit application has been received by the New Orleans District of the U.S. Army Corps of Engineers pursuant to: [X] Section 10 of the Rivers and Harbors Act of March 3, 1899 (30 Stat. 1151; 33 USC 403); and/or [ ] Section 404 of the Clean Water Act (86 Stat. 816; 33 USC 1344).

BARGE MOORING ON MISSISSIPPI RIVER IN ORLEANS PARISH

NAME OF APPLICANT: Harbor Towing and Fleeting c/o Lanier & Associates Engineers at 4101 Magazine Street, New Orleans, Louisiana 70115.

LOCATION OF WORK: On the right descending bank of the Mississippi River, approximately 90.5 miles above Head of Passes, New Orleans, in Orleans Parish, Louisiana, as shown on the attached drawings.

CHARACTER OF WORK: To install and maintain six monopile breasting dolphins to moor five tiers of barges eight wide. The new barges will extend approximately 400 feet from the shoreline and in line with the existing barges at the site. This is an expansion to an existing barge mooring operation previously permitted under the same number on October 24, 1997. No excavation or fill will occur; no compensatory mitigation is anticipated at this time.

The comment period for the Department of the Army Permit will close **30 days** from the date of this public notice. Written comments, including suggestions for modifications or objections to the proposed work, stating reasons therefore, are being solicited from anyone having interest in this permit request. Letters must reference the applicant's name and the subject number, be addressed and mailed to the above address, ATTENTION: REGULATORY BRANCH.

New Orleans Wetlands

How the Gulf Restoration Network uses these

1. Read the public notices regularly.
2. Identify applications for inappropriate things (like shopping malls).
3. Contact

In the past, Scott has had to do this manually. But he doesn't really have time for that.

We're using a computer program to make the first two of these steps easier.

New Orleans Wetlands

My script extracts this information.

May 7, 2012

United States Army
Corps of Engineers
New Orleans District
Regulatory Branch
Post Office Box 60267
New Orleans, Louisiana 70160-0267

(504) 862-2225
Project Manager
Brad LaBorde
Permit Application Number
MVN 2012-1000 EOO

State of Louisiana
Department of Environmental Quality
ATTN: Water Quality Certifications
Post Office Box 4313
Baton Rouge, Louisiana 70821-4313

(225) 219-3225
Project Manager
Jamie Phillippe
WQC Application Number
WQC 120507-01

Interested parties are hereby notified that a permit application has been received by the New Orleans District of the U.S. Army Corps of Engineers pursuant to: [X] Section 10 of the Rivers and Harbors Act of March 3, 1899 (30 Stat. 1151; 33 USC 403); and/or [X] Section 404 of the Clean Water Act (86 Stat. 816; 33 USC 1344).

Application has also been made to the Louisiana Department of Environmental Quality, Water Quality Certifications, for a Water Quality Certification (WQC) in accordance with statutory authority contained in LRS30:2047 A(3), and provisions of Section 401 of the Clean Water Act (P.L.95-17).

LAUNCH BOAT LANDING WITHIN THE MISSISSIPPI RIVER

NAME OF APPLICANT: Belle Chasse Marine Transportation, Inc., % Richard Wright and Associates, Inc., 1013 Colony Place, Metairie, Louisiana, 70003.

LOCATION OF WORK: In and adjacent to the Mississippi River on the left descending bank, at Mississippi River mile 126.4 above Head of Passes, near Norco, in Jefferson Parish, Louisiana, latitude: 29.99639/longitude: -90.41444, as shown on the enclosed drawings.

CHARACTER OF WORK: Install and maintain a landing barge, hinged ramp, and associated structures for a launch boat landing to safely transfer personnel to/from marine vessels. A 48 square foot concrete slab will be placed on the batture with a 4-foot wide, 120-foot long hinged ramp extending towards the river for access to a 20-foot wide, 40-foot long landing barge. Clearing of 0.01 acres of Mississippi River batture is proposed for slab and ramp construction. No compensatory mitigation measures have been proposed for this project at this time.

The comment period for the Department of the Army Permit and the Louisiana Department of Environmental Quality WQC will close **30 days** from the date of this joint public notice. Written comments,

New Orleans Wetlands

It also

- runs automatically every day
- saves all of the files
- checks for changes in files
- produces a spreadsheet of the extracted information
- hosts all of this on a website that Scott can access

We're still working out the kinks, but the initial goal is that Scott will be able to use the spreadsheet to quickly find notices that he should look into further. Then he'll read the notice and take whatever actions make sense.

Financial products in Africa

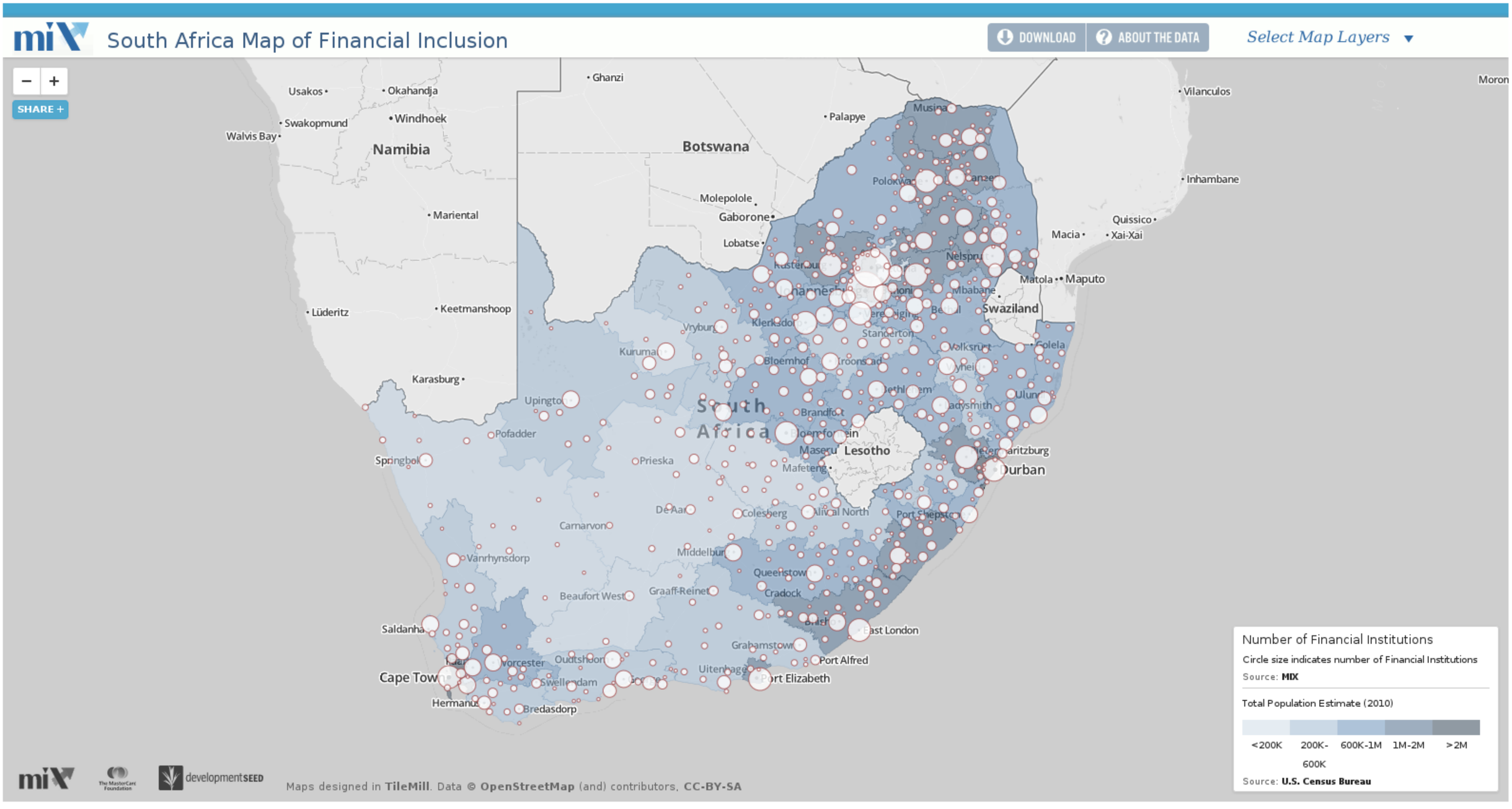
Measuring access to financial products

- The Microfinance Information Exchange

> *MIX delivers data services, analysis, research and business information on the institutions that provide financial services to the world's poor.*

- Who has access to what financial products?
- I scraped locations of all sorts of credit providers

Map of the financial sector in South Africa



Ergonomics research

Ergonomics research

In school, I studied how people use computers and toilets (not at the same time)

Scripts for

- Tidying data
- Running models
- Plotting data

Ergonomics research

Sample for a questionnaire

- I was studying the postures in which people use toilets
- I wanted a sample of students
- My university had a public database of all student, faculty and alumni email addresses.

Frivolity

Middle names

While selecting the sample for that toilet study, I started wondering how many people have middle names.

I asked the US Census.

Subject

What proportion of people have middle initials?

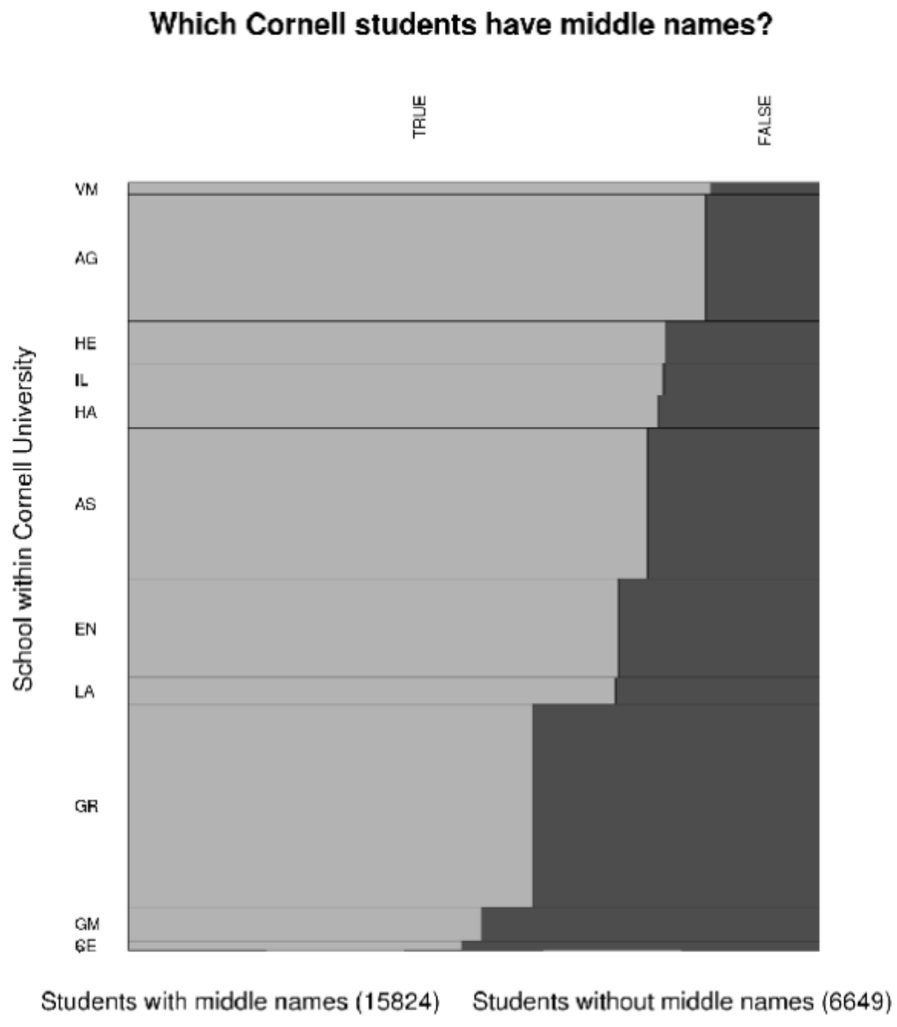
Discussion Thread

Response Via Email(CLMSO - EMM) - 03/14/2011 16:04
Thank you for using the US Census Bureau's Question and Answer Center. Unfortunately, the subject you asked about is not one for which the Census Bureau collects data. We are sorry we were not able to assist you.

(http://blog.scraperwiki.com/2012/06/15/middle-names-in-the-united-states-over-time/)

Middle names

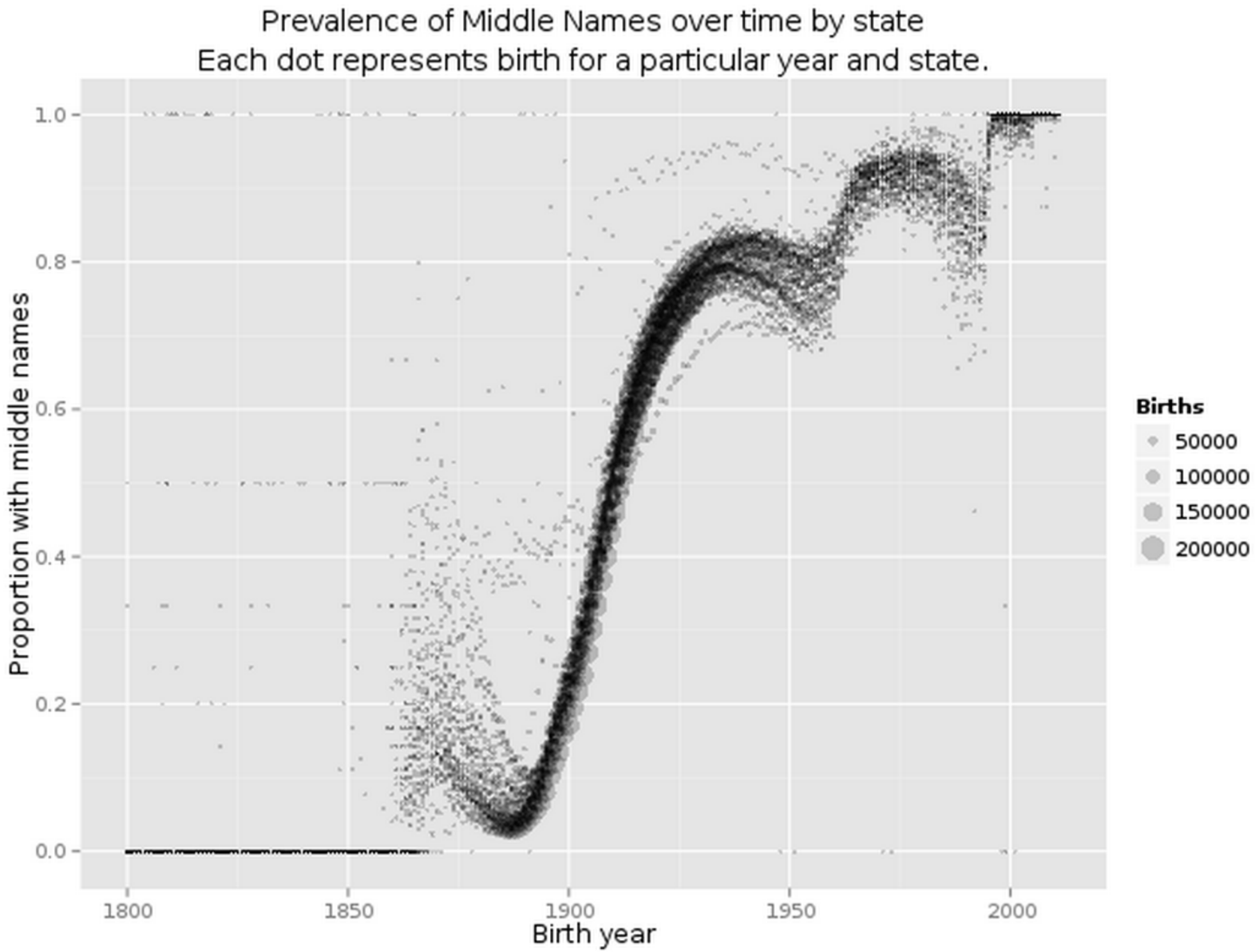
The Census couldn't tell me, so I looked at that university database



Two-thirds had middle names.

Middle names

A few months ago, I looked at this again with all dead Americans



End of bragging-about-projects

What is the point?

- I use numbers to learn about how people work.
- If I had clean data, the analysis would be easy.
- Cleaning the data is most of the work of an analysis.
- I know enough about computers to do all this, but you shouldn't need to.

My actual work

ScraperWiki

We're making a platform to make data projects easier for you.

- Simplify computer infrastructure so you don't have to learn it.
- Automatically provide obvious tools: scheduling, a web API, &c.
- Make projects more visible.

And you can brag too

How do you learn about this data stuff?

Find something that annoys you and for which no good tools exist.

Try making something to make the thing less annoying.

- Start with a small part of that task.
- Unsophisticated programs can get you pretty far.
- Use existing software libraries so you don't have to write your own

If that's still too hard, do something else to make the thing less annoying.

Thomas Levine
ScraperWiki

Learn to Scrape

Cheatsheet

Follow along here.

Python | Ruby

```python
from urllib2 import urlopen
from lxml.html import fromstring
from scraperwiki.sqlite import save

# Load
html = urlopen('http://example.com').read()
x = fromstring(html)

# Select a table
table = x.cssselect('table')[2]
tr = table.cssselect('tr')[7]

# Select links
links = x.cssselect('a')
print [a.attrib['href'] for a in links]

# Combine into a dictionary
header = ['foo', 'bar', 'baz']
cell_content = [td.text_content() for td in tr.cssselect('td')]
data = dict(zip(header, cell_content)

# Save to ScraperWiki's datastore
save([], data)
```

Open these pages

- scraperwiki.thomaslevine.com (these slides)
- scraper page

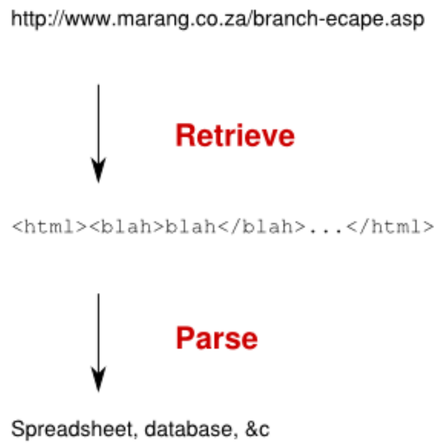
Agenda

1. Introduction to scraping
2. Introduction to ScraperWiki
3. Collective bargaining agreements scraper
4. Analyzing the scraped data

A computer can do anything an intern can.

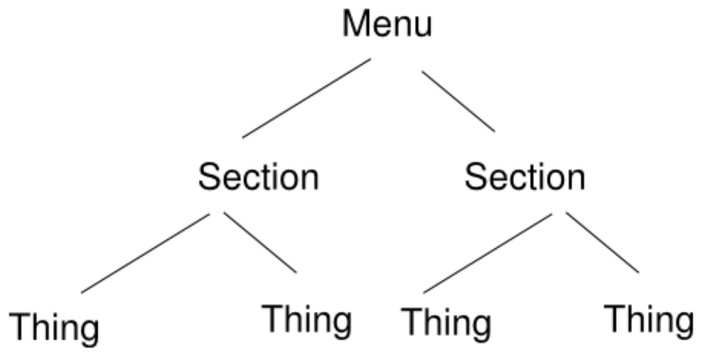
"Scraping"

- "Scraping" involves retrieving some raw document and parsing it to turn it into something else.
- Today, we're particularly concerned about raw documents that are websites.
- The retrieval and parsing are convenient places to divide the script.



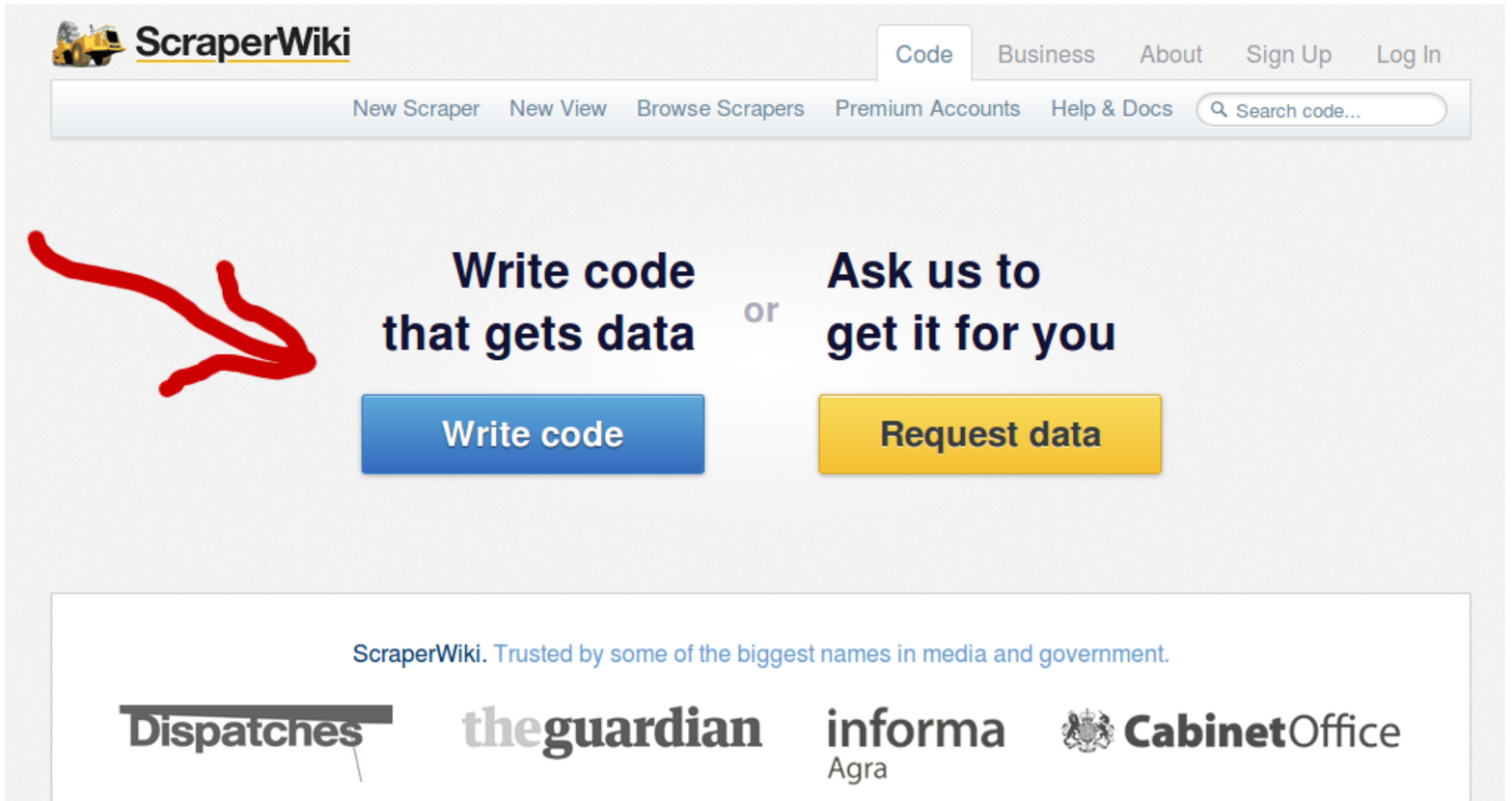
Hierarchy

Parsing one document may lead you to other documents to retrieve.



Using ScraperWiki: Starting a script.

Sign up, then click this.



Follow along with me.

Follow along here.

- You'll see what I type.
- If you want to change something and see what happens, click "COPY".

Hello World

Python | Ruby

```
print("Hello World")
```

Hello Web

Python | Ruby

```
download=urlopen("http://newshackdaysf.tumblr.com/")
print(download.read()) #This method is annoying; see below.
print(download.read())
```

Hello Datastore

Python | Ruby

```
data={
  "firstname":"Joseph",
  "lastname":"Pulitzer",
  "birthday":"1847-04-10"
}
save([],data)
```

Try saving something else.

Scraping project for the course

The Department of Labor collective bargaining agreement filings website

Pages types

- Main page with links to tables
- Many filings per table
- One pdf per filing

Let's start coding

Outline of the scraper

1. Download the page.
2. Select the table.
3. Select the table rows from the table.
4. Select the table cells from the table row.
5. Print the table cells.

Import one of the tables

1. Print this one: http://www.dol.gov/olms/regs/compliance/cba/Cba_CaCn.htm"
2. Load the html with fromstring. This gives us more power, as you'll see later.
3. Traverse the table.

HTML Tables

```
<table>
  <tr>
    <th>Location</th>
    <th>Union</th>
  </tr>
  <tr>
    <td>AL</td>
    <td>BBF</td>
  </tr>
  <tr>
    <td>NM</td>
    <td>SPGU</td>
  </tr>
  <tr>
    <td>NE</td>
    <td>BBF</td>
  </tr>
</table>
```

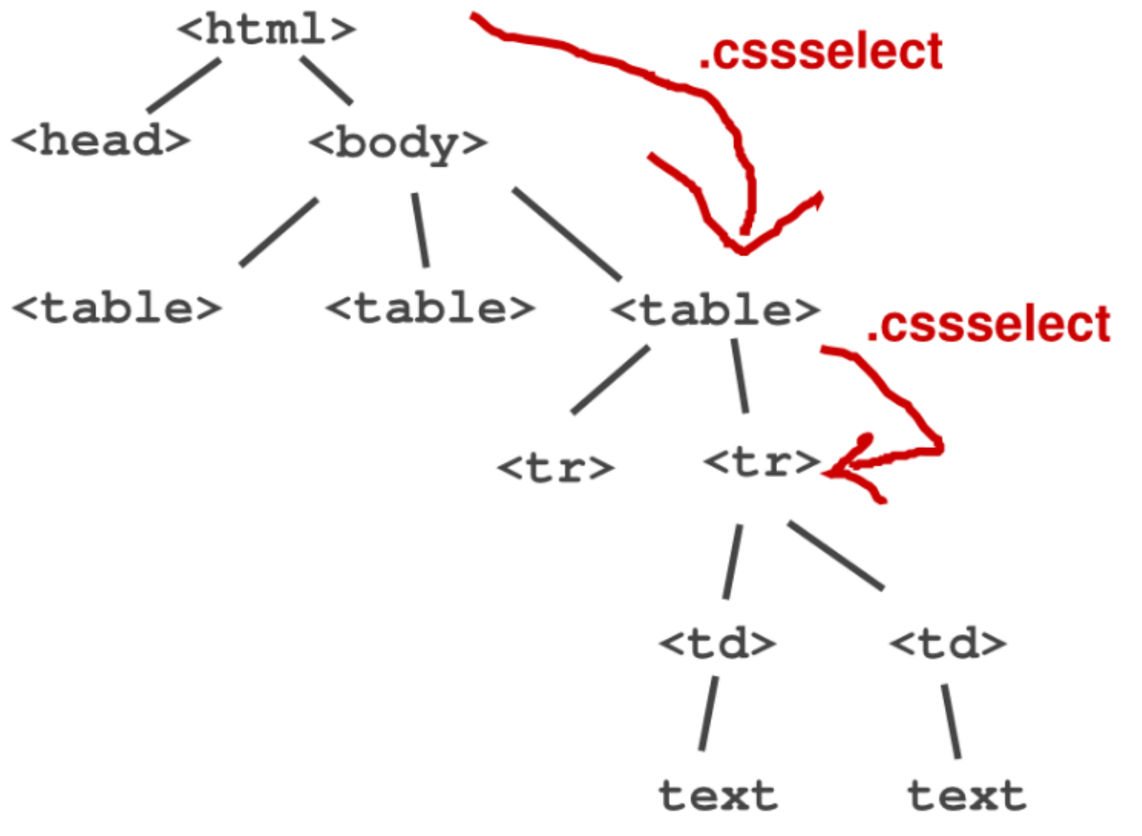
Location	Union
AL	BBF
NM	SPGU
NE	BBF

Selecting the table:

1. Look for the table in the html by searching for the first row.
2. How can we use CSS to get the table?
3. Select tables with lxml, and print it with tostring
4. Select the table of interest with the list index.

Deeper selections

1. Select a table row from the table, and print it.
2. Select a table cell from the table row, and print it.
3. Get the text out of a table cell with .text_content()



Iterate

1. For each cell in a row, print the cell.
2. For each cell in a row, print the plain text content of the cell.
3. For each row in the table, for each cell in a row, print the text content.

Break

Review

Done

1. Download the page.
2. Select the table.
3. Select the table rows from the table.
4. Select the table cells from the table row.
5. Print the table cells.

To do

1. Save the data instead of printing them.
2. Clean up the data before saving them.
3. Run on all of the tables.
4. Do something with the pdfs?

Row-level data

We are currently just printing individual cell data. We would like to save data by agreement (by row).

Dictionary

The save function wants a dictionary.

We have this.

Python Ruby

```
#List of column names
[
   'employer','download','location','union',
   'local', 'naics', 'num_workers', 'expiration_date'
]

#List of row values
[
   'California  Processors Inc.\r\n    ', 'Not Available', 'CA', 'IBT',
   '748,857,601', '311421', '11000', '6-30-06', '123', '422'
]
```

We want it to look more like this.

Python Ruby

```
{
   'download': 'Not Available', 'employer': 'California  Processors Inc.\r\n    ',
   'expiration_date': '6-30-06', 'local': '748,857,601',
   'location': 'CA', 'naics': '311421',
   'num_workers': '11000', 'union': 'IBT'
}
```

Combine the header with the data row.

Once we've combined them, we have a working scraper; look at the scraper overview page.

What we'se done so far

The Department of Labor collective bargaining agreement filings website

Pages types

- Main page with links to tables
- Many filings per table <- We just did this.
- One pdf per filing

Removing the header row

1. We saved the header row to the datastore; we don't want that.
2. The .pop method

Cleaning the data

A few messy columns

1. State
2. Number of workers
3. Date
4. Local
5. PDF links

More cleaning

Columns with counts

1. You can't add strings.
2. Convert to integer.

Date

1. We want to change to ISO format (YYYY-MM-DD)
2. strptime
3. This checks that dates are dates
4. This also allows us to add dates and stuff.

State

1. We want to get the two-letter code out of the location column.
2. Regular expressions

PDF links

1. Select a tags
2. Get the href attributes with .attrib

Verifying states

Here's a list of state codes.

```
[
    'AL', 'AK', 'AZ', 'AR', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA',
    'HI', 'ID', 'IL', 'IN', 'IA', 'KS', 'KY', 'LA', 'ME', 'MD',
    'MA', 'MI', 'MN', 'MS', 'MO', 'MT', 'NE', 'NV', 'NH', 'NJ',
    'NM', 'NY', 'NC', 'ND', 'OH', 'OK', 'OR', 'PA', 'RI', 'SC',
    'SD', 'TN', 'TX', 'UT', 'VT', 'VA', 'WA', 'WV', 'WI', 'WY'
]
```

Multiple pages

1. Here is a list of urls. Try running the script for each url in the list.

```
[ "http://www.dol.gov/olms/regs/compliance/cba/Cba_CaCn.htm",
  "http://www.dol.gov/olms/regs/compliance/cba/Cbau_mamh.htm",
  "http://www.dol.gov/olms/regs/compliance/cba/Cba_NfOz.htm" ]
```

2. Where did that list come from?
3. Scraping the main page
    1. Downloading the page
    2. Load into lxml
    3. Select the links (a tags)
    4. Sort the links by public/private
    5. Save public/private to the datastore
4. Loop over this list.

The final scraper is here.

Now you can analyze

- Aggregate by state
- Find out-lying collective bargaining agreements
- Group by categories
- More ideas?

Review

- Computers can do anything an intern can.
- Downloading web pages
- Parsing an html table
- Validating data types
- This allows new analysis