**Downloading Web Data Without Scraping**

**Michelle Minkoff and Scott Klein, ONA Atlanta 2013**

**Software to Install**

**Google Chrome**

Internet Explorer and Firefox both have excellent developer tools but the Web Inspector in Chrome will be basis of our examples. Some examples use Firefox extensions, but if you don’t have it, no worries.

JSONView Chrome Extension: http://goo.gl/njpwwh

**CSVKit**

<http://csvkit.readthedocs.org/en/latest/index.html>

Requires Python. Mac/Linux comes with Python installed. Use Anthony DeBarros’s great guide to install Python on Windows. Install Python 2.7.x because CSVKit doesn’t work with Python 3. http://goo.gl/Nm90gD

**Google Spreadsheets**

You can import an HTML table directly by typing =ImportHTML(“url”, “elementtype”, numberElement on page)

Ex: =ImportHTML(“http://www.atlanta.k12.ga.us/Page/832”, “table”, 2)

**Tools we’re using**

**Open Refine**

<http://openrefine.org/>

Tool to inspect and manipulate spreadsheet files, allowing to run queries on it to manipulate it the way you would like.

**Scraper**

<https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd?hl=en>

Click on a type of information (names, emails, URLs, etc, and then right-click (control-click) and Scrape Similar. There’s an option to bring the results into a Google spreadsheet.

**Tabula**

<http://tabula.nerdpower.org/>

Tabula turns tabular PDF data into tables. Free software from the Knight-Mozilla OpenNews project.

**DownThemAll**

<https://addons.mozilla.org/en-US/firefox/addon/downthemall/>

A Firefox extension that detects what types of assets you might want to download on a page and allows you to download them, well, all. Doesn’t work for everything, but a good quick one to try.

**OutwitHub**

<https://addons.mozilla.org/en-us/firefox/addon/outwit-hub/>

Break your page down into its elements that it is made up of. Grab all pictures on a page at once, create a scraper using various HTML elements as start and end points. This is quite powerful, if you learn how to use it well. More robust than DownThemAll.

**JQ**

<http://stedolan.github.io/jq/download/>

Helpful command line utility for working with JSON files.

**JSON Viewer**

<http://jsonviewer.stack.hu/>

**Other Important Links**

**Dan Nguyen’s Terrific Web Inspector Guide**

http://dannguyen.github.io/NICAR/2012/02/25/nicar-2012-inspect-the-web-with-your-browsers-web-inspector/

**When grabbing data, please remember to make sure you have permission to get the information you are grabbing. Just because you can take photos off a site, doesn’t mean you can use them for your project. With great power comes great responsibility!**

**Example sites we’re using during the workshop**

1. Atlanta school clusters – for html table -- <http://www.atlanta.k12.ga.us/Page/832>,
2. Atlanta food and drug officials page – for Scraper - <http://dslo.afdo.org/results/?q=Georgia&unifyfda=1&bystate=1&selected_facets=area_exact:%22100%22>
3. Atlanta image search - <https://www.google.com/search?q=atlanta&espv=210&es_sm=119&source=lnms&tbm=isch&sa=X&ei=PHtcUrakJZT54AOJyYDoCg&ved=0CAcQ_AUoAQ&biw=1449&bih=1102>
4. <http://projects.propublica.org/recovery/>
5. <http://projects.propublica.org/graphics/intern-suits>
6. <http://transform.mo.gov/map/>
7. JSONView