



## CodeActually

Dr. Cindy Royal

Texas State University

School of Journalism and Mass Communication

---

### Basic Web Scraping

Scraping is a process by which you can extract data from an html page or pdf, into a CSV or other format, so you can work with it in Excel or another spreadsheet and use in your visualizations. Sometimes you can just copy and paste data from an html table into a spreadsheet, and all will be fine. Other times, not. So, you have to create a computer program that allows you to identify and scrape the data you want. There are numerous ways in which you can do this, some that involve programming and some that don't.

### Scraping with Google Spreadsheet

If copying and pasting directly into a spreadsheet doesn't work, you can try using Google Spreadsheet functions to scrape the data. Open a new Google Spreadsheet. We are going to scrape the data from the Texas Music Office site. The url <http://governor.state.tx.us/music/musicians/talent/talent/>, goes to the first page of the directory, listing the bands that start with A.

In the first cell of your spreadsheet, type the function:

```
=ImportHtml("http://governor.state.tx.us/music/musicians/talent/talent/", "table", 1)
```

- First argument is the url
- Second argument tells it to look for a table (the other element allowed here is "list"),
- Third argument is the index of the table (if there are multiple tables on the page).

You will have to look at the html to find the table from which you are trying to get the data or through trial and error by changing the third number.

Give it a couple seconds and you should see the data directly from the table in your spreadsheet. Easy!

Google Spreadsheets has a few other functions that can be helpful in scraping data.

- ImportFeed will scrape from an RSS feed. Try:  

```
=ImportFeed(http://news.google.com/news?pz=1&cf=all&ned=us&
```

hl=en&topic=t&output=rss)

Find any RSS feed by looking for the RSS icon. This link pulls current items from Google News.

- If your data is already in CSV format, you can save a step to bring it into your spreadsheet by using ImportData. This will also scrape the data directly from the site, so it will pull from the most recent version of the file. Try:

```
=importData("http://www.census.gov/popest/data/national/totals/2012/files/NST_EST2012_ALLDATA.csv")
```

Of course you can also just open the csv in the spreadsheet program and follow the instructions for importing it.

### **Chrome Scraper Extension**

<https://chrome.google.com/extensions/detail/mbigbapnjcgafohmbkdlecaccepngjd> - free extension for Chrome. Select content on a page, use the context menu to Scrape Similar. Can export results to a Google Doc.

Download and install the Chrome Scraper extension. Go to this page:  
[http://en.wikipedia.org/wiki/List\\_of\\_Academy\\_Award-winning\\_films](http://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films)

It includes a list of Academy Award-winning Films in a table. Select the first row of the table. Ctrl-click and choose Scrape Similar. You should see the entire table. You can easily Export to Google Docs with the button.

The screenshot shows a browser window with the URL `ia.org/wiki/List_of_Academy_Award-winning_films`. A scraper tool is active, displaying a table of data extracted from the page. The scraper's selector is `//div[4]/table[1]/tbody/tr[td]`. The table lists 21 films with columns for Film, Year, Awards, and Nominations.

	Film	Year	Awards	Nominations
1	20,000 Leagues Under the Sea	1954	2	3
2	2001: A Space Odyssey	1968	1	4
3	7 Faces of Dr. Lao	1964	0 (1)	1
4	7th Heaven	1927/28	3	5
5	8 Mile	2002	1	1
6	Abyss, TheThe Abyss	1989	1	4
7	Accidental Tourist, TheThe Accidental Tourist	1988	1	4
8	Accountant, TheThe Accountant	2001	1	1
9	Accused, TheThe Accused	1988	1	1
10	Adaptation.	2002	1	4
11	Adventures of Don Juan	1949	1	2
12	Adventures of Priscilla, Queen of the Desert, TheThe Adventures of Priscilla, Queen of the Desert	1994	1	1
13	Adventures of Robin Hood, TheThe Adventures of Robin Hood	1938	3	4
14	Affliction	1998	1	2
15	African Queen, TheThe African Queen	1951	1	4
16	Age of Innocence, TheThe Age of Innocence	1993	1	5
17	Air Force	1943	1	4
18	Airport	1970	1	10
19	Aladdin	1992	2	5
20	Alamo, TheThe Alamo	1960	1	7
21	Alaskan Eskimo, TheThe Alaskan Eskimo	1953	1	1

Notice the XPath description. This is the code the scraper used to find the table.

```
//div[4]/table[1]/tbody/tr[td]
```

It found the first table in the fourth div and extracted elements in the tds.

This method did not get the links. To do that, let's try just right clicking on the first item in the table (unselect the entire row). This should find all the links. Take a look at the difference in XPath code:

```
//td/i/a
```

This technique finds all the links within `<i>` tags in the tds.

The screenshot shows a web browser window with a scraper tool overlaid on a Wikipedia page. The scraper tool is titled "Scraper - List of Academy Award-winning films - Wikipedia, the free encyclopedia". It has a "Selector" field containing the XPath expression `//td//i/a` and a "Columns" section with two columns: "Link" and "URL". The "URL" column is selected. The scraper tool is displaying a list of 19 film titles and their corresponding URLs. The browser window shows the Wikipedia page with a table of film titles and a sidebar with navigation links.

	Link	URL
1	20,000 Leagues Under the Sea	/wiki/20,000_Leagues_Under_the_Sea_(1954_film)
2	2001: A Space Odyssey	/wiki/2001:_A_Space_Odyssey_(film)
3	7 Faces of Dr. Lao	/wiki/7_Faces_of_Dr._Lao
4	7th Heaven	/wiki/Seventh_Heaven_(1927_film)
5	8 Mile	/wiki/8_Mile_(film)
6	Adaptation.	/wiki/Adaptation_(film)
7	Adventures of Don Juan	/wiki/Adventures_of_Don_Juan
8	Affliction	/wiki/Affliction_(film)
9	Air Force	/wiki/Air_Force_(film)
10	Airport	/wiki/Airport_(1970_film)
11	Aladdin	/wiki/Aladdin_(1992_Disney_film)
12	Albert Schweitzer	/wiki/Albert_Schweitzer_(film)
13	Alexander's Ragtime Band	/wiki/Alexander%27s_Ragtime_Band_(film)
14	Alice in Wonderland	/wiki/Alice_in_Wonderland_(2010_film)
15	Alice Doesn't Live Here Anymore	/wiki/Alice_Doesn't_Live_Here_Anymore
16	Alien	/wiki/Alien_(film)
17	Aliens	/wiki/Aliens_(film)
18	All About My Mother	/wiki/All_About_My_Mother
19	All That Jazz	/wiki/All_That_Jazz_(film)

You can learn more about XPath syntax at [http://www.w3schools.com/xpath/xpath\\_syntax.asp](http://www.w3schools.com/xpath/xpath_syntax.asp).

## More Scraping Tools and Resources

There are many other tools that can be used effectively to scrape data from Web pages. Here are a few additional resources:

- OutWit Hub – a Firefox extension and Desktop program that can provide some advanced scraping capabilities
- ProPublica's Scraping for Journalism: A Guide For Collecting Data - <http://www.propublica.org/nerds/item/doc-dollars-guides-collecting-the-data>
- Getting to Grips with ScraperWiki For Those Who Don't Code - <http://datamineruk.wordpress.com/2011/07/21/getting-to-grips-with-scraperwiki-for-those-who-dont-code/>
- Web Scraping for Non-Programmers by Michelle Minkoff - <http://michelleminkoff.com/outwit-needlebase-hands-on-lab/>