**Grab data from the web (without programming)  
NICAR ‘16 - Denver - March 10-13**Robert Gebeloff - [rgebeloff@nytimes.com](mailto:rgebeloff@nytimes.com)   
Tom Johnson — [tom@jtjohnson.com](mailto:tom@jtjohnson.com)

This document at <http://bit.ly/NICAR-webscraping>

**Online Tools**

* [**Web Scraper**](https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehllkliplmbmhn?hl=en)   
  *“Web Scraper is a chrome browser extension built for data extraction from web pages. Using this extension you can create a plan (sitemap) how a web site should be traversed and what should be extracted. Using these sitemaps the Web Scraper will navigate the site accordingly and extract all data. Scraped data later can be exported as CSV.”*
* [**Import.io**](https://import.io/)*“...free desktop app to help you scrap all the data you need from an unlimited amount of web pages. The service treats each page as a potential data source to generate API from. If the page you’ve submitted has been previously processed, you can access its API and get some of the data.”*
* ***#1 demo — HTML Table Plugins***  *-- Both* [*Chrome*](https://chrome.google.com/webstore/detail/table-capture/) *and* [*Firefox*](https://addons.mozilla.org/en-US/firefox/addon/dafizilla-table2clipboard/) *have plugins [FOR EXAMPLE???] that allow you to easily copy and paste data tables from Web sites into excel.*
* ***Down Them All! --*** *Another handy Firefox* [*plugin*](https://addons.mozilla.org/en-US/firefox/addon/downthemall/) *that allows you to bulk-download files from a page. Includes simpler filtering tools so you can, for example, download just the .xls files or just the .zip files with the string “county” in the name (\*county\*.zip).*
* ***WGET*** *-- An oldy but goodie, a command line tool for bulk downloading files. Let’s say you want to scrape state data from a site and you know that each state has its own page with the identical URL but the state FIPS code as a suffix, as in http://sitename.com/state/34 and* [*http://sitename.com/state/36*](http://sitename.com/state/36)*. You could use excel to make a list of all the possible urls, save the list in a text file, and wget -i list.txt to grab all 50 files. Or you can even mirror and entire Web site like “wget --mirror –w 2 –p --html-extension –-convert-links –P /Users/robertgebeloff/Downloads http://www.sitetocopy.com “*
* ***Mirror An Entire Site --*** *Something like this* [*https://www.httrack.com/*](https://www.httrack.com/)*, not sure what the state of the art is these days.*
* *Use* ***Selenium*** *to turn your browser into a bot. Here's a good starting point* [*http://thiagomarzagao.com/2013/11/12/webscraping-with-selenium-part-1/*](http://thiagomarzagao.com/2013/11/12/webscraping-with-selenium-part-1/)

# [*Selenium Tutorial: Web Scraping with Selenium and Python*](http://www.marinamele.com/selenium-tutorial-web-scraping-with-selenium-and-python)

**Online Tutorials**

* [**An XML miracle (or how to find a Trader Joe’s that sells beer)**](http://datadrivenjournalism.net/news_and_analysis/an_xml_miracle_or_how_to_find_a_trader_joes_that_sells_beer#sthash.Eqiii5FP.dpuf)*“Look under the hood at data structures we encounter on the Web. [This] explains how to find underlying data sources that are not immediately obvious, and how these sources are structured – in many cases, a Web page will incorporate data in a format known as XML.”*
* [**Caching Tutorial for Web Authors and Webmasters**](https://www.mnot.net/cache_docs/)What’s a Web Cache? Why do people use them?  
  A Web cache sits between one or more Web servers (also known as origin servers) and a client or many clients, and watches requests come by, saving copies of the responses — like HTML pages, images and files (collectively known as representations) — for itself. Then, if there is another request for the same URL, it can use the response that it has, instead of asking the origin server for it again.  
  There are two main reasons that Web caches are used:
  + To reduce latency — Because the request is satisfied from the cache (which is closer to the client) instead of the origin server, it takes less time for it to get the representation and display it. This makes the Web seem more responsive.
  + To reduce network traffic — Because representations are reused, it reduces the amount of bandwidth used by a client. This saves money if the client is paying for traffic, and keeps their bandwidth requirements lower and more manageable.
  + [How to Cache a Website on Google](http://smallbusiness.chron.com/cache-website-google-46565.html) by Marissa Robert,

**Related Tip-sheets**

* **CATEGORY:** [**SCRAPING TIPS**](https://learn.scrapehero.com/category/scraping-tips/?subscribe=success#blog_subscription-2)
  + Scalable do-it-yourself scraping – How to build and run scrapers on a large scale
  + Why \*not\* scrape yourself
  + How to prevent getting blacklisted while scraping [and more]
* [**A Guide to Web Scraping Tools**](http://www.garethjames.net/a-guide-to-web-scrapping-tools/) **by Gareth** [DECEMBER 5, 2014] A bit dated, but this will still point you to some resources that have evolved.
* [**HOW TO PREVENT GETTING BLACKLISTED WHILE SCRAPING**](https://learn.scrapehero.com/how-to-prevent-getting-blacklisted-while-scraping/)

**Articles and Books**

* [**The Ultimate Guide to Web Scraping**](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&cad=rja&uact=8&ved=0ahUKEwiY1eiKvNfKAhVD9WMKHTO9A5EQFghgMAg&url=https%3A%2F%2Fblog.hartleybrody.com%2Fweb-scraping-guide%2F&usg=AFQjCNGOPU4cAA0C-MaWDRgnxYVPnCzcrg&sig2=0cp1w_Ty2O8OpO5LetikTQ)*“No prior knowledge of web scraping is necessary to follow along — the book is designed to walk you from beginner to expert, honing your skills and helping you become a master craftsman in the art of web scraping.”*
* [**Wikipedia - Web Scraping**](https://en.wikipedia.org/wiki/Web_scraping)Includes links to 27 “Notable Tools”
* [**The Top Pitfalls of Web Scraping**](http://www.connotate.com/top-pitfalls-web-scraping/) by Vincent Sgro
* [**I Don’t Need No Stinking API: Web Scraping For Fun and Profit**](https://blog.hartleybrody.com/web-scraping/)*“If you’ve ever needed to pull data from a third party website, chances are you started by checking to see if they had an official API. But did you know that there’s a source of structured data that virtually every website on the internet supports automatically, by default?”* [Also see his book, “[The Ultimate Guide to Web Scraping](https://www.simplegoods.co/i/QNZWMKHL),” a good intro.]
* “Scraping - Beyond the Basics” in “[**Data Wrangling Handbook**](https://datapatterns.readthedocs.org/en/latest/index.html#)” — Especially good for those first entering the galaxy of data scraping and analysis.
* [**A beginner’s guide to data scraping in Python**](http://www.eco.utexas.edu/.../Lay/Python.docx) by Garrett Lay  
  Good, step-by-step instructions that also includes “Scraping with Chrome,” — “a very powerful tool ... that allows scraping basic data in an easier and much friendlier fashion when compared to using a python environment like Canopy.”

**Misc. Tools, etc. we find handy**

* [**PDFTables.com**](http://pdftables.com)  
  *“This is the first self-service, web-based product designed for getting volumes of data from PDFs. It’s super fast to convert individual PDFs, and there’s a* [*web API*](http://pdftables.com/api) *to automate more.”*
* [*Guide to Dealing With PDFs*](https://github.com/gebelo/gijc/blob/master/pdf_wrangling.docx) *-- This handout walks you through a myriad of options for dealing with data trapped inside PDFs, from free to expensive software to programming techniques.*
* **Let’s Git Started: A Beginner’s Guide to Version Control Software** - See more at: <http://blog.flatironschool.com/lets-git-started-a-beginners-guide-to-version-control-software/#sthash.y9P3DLfJ.dpuf>
* **Version control concepts and best practices**[**https://homes.cs.washington.edu/~mernst/advice/version-control.html**](https://homes.cs.washington.edu/~mernst/advice/version-control.html)*“This document is a brief introduction to version control. After reading it, you will be prepared to perform simple tasks using a version control system, and to learn more from other documents that may lack a high-level conceptual overview. Most of its advice is applicable to all version control systems, but its examples use* [*Git*](http://git-scm.com/)*,* [*Mercurial (Hg)*](http://mercurial.selenic.com/)*, and* [*Subversion (SVN)*](http://subversion.apache.org/) *for concreteness.”*