CHAPTER III

# Scraping Data

“Website scraping is a dark art.”[[1]](#footnote-1)

**What is data scraping?**

Data scraping is a technique that allows you to get data that is usable and user-friendly. By scraping, you take the data from a human-readable source (i.e a .pdf document or webpage) and make it readable to a machine (i.e a .csv file that can be read by Tableau or R).

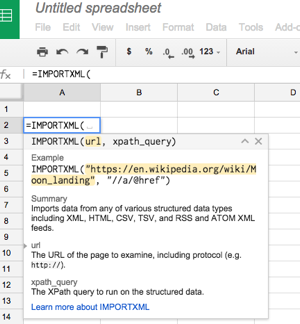
**Cardinal Rules of Data-Scraping**

1. Do not scrape if you do not have to: always look for a copy of the dataset in some format that you can easily convert to a .csv file using Excel or other conversion tool (import.io). Chances are that the data is out there in some form that will make your life easier.

2. Look for an option to download a .csv file directly from the source. Most of the sites that carry information about development have ways to guide readers to build their own report and download the data that they need.

**Design the scraper**

*What type of file do you have?*



*A .csv file: you are done* (go to the data format section).

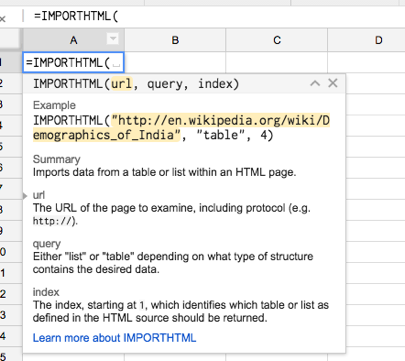
*An .xml file*: use a converter tool. You can also use google spreadsheets command =importxml

To use it you will need to be familiar with XPATHS[[2]](#footnote-2) (they are used to navigate through the elements of an .xml document). XML documents are trees of nodes. This can be thought of as similar to a family tree with various connections and relationships. For example:

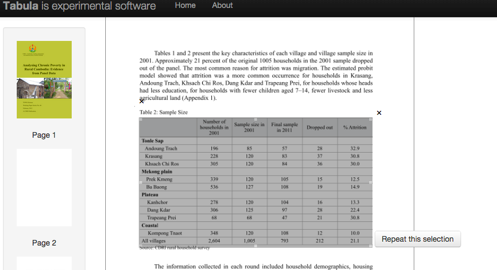
// development: means get all the nodes that are called “development”

@lang=eng: means get all the nodes that have the attribute of the English language.

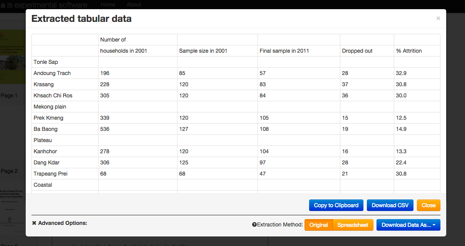
*A .json file:* use a converter tool such as <http://www.convertcsv.com/json-to-csv.htm> (there are many others out there that can be found through a Google search). You have to be careful because .json files do not always fit into the .csv table format.



*A html table:* use the =importhtml command in google. You can export a table or a list imbedded in a html document by writing the url, the type of element you want (table or list) and the number of the table or list.



*A table in a .pdf file*: use Tabula. You can download Tabula at tabula.technology and follow the download instructions. Once you have installed Tabula, just open it, upload your .pdf document and manually select the table you need to export.



Once you have selected your data, wait for Tabula to extract it and download as .csv:

**APIs**

*What’s an API?*

An Application Programming Interface (API) is a software intermediary that makes it possible for application programs to interact with each other and share data.[[3]](#footnote-3) Each API has its own unique instructions to access data and will return a unique set of data structures when data is requested. An API can be used by applications to access data sets in machine-readable formats like csv, xml, and json. Some examples are:

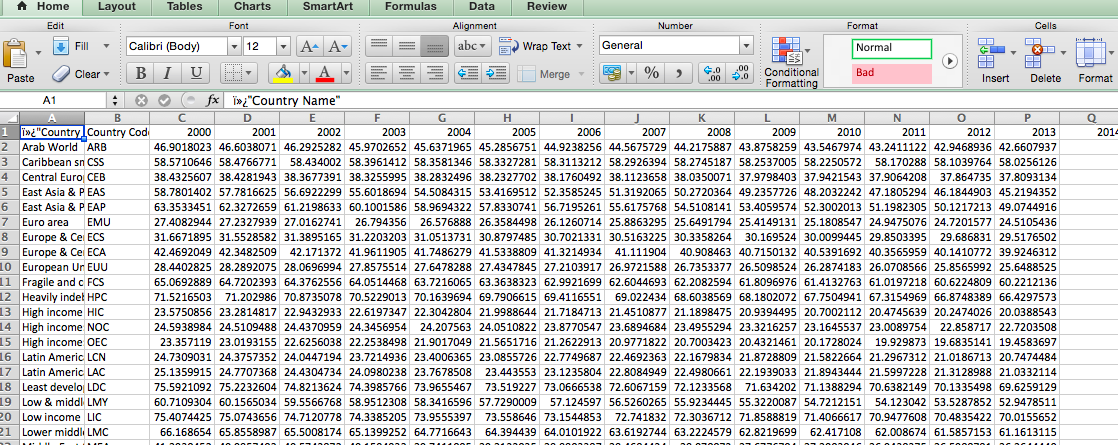
World Bank’s data API: <http://data.worldbank.org/node/9>

Twitter’s search API: <https://dev.twitter.com/overview/api>

API included in most CKAN datastores: <http://docs.ckan.org/en/latest/api/index.html>

Example: go to [http://api.worldbank.org/countries/all/indicators/SP.RUR.TOTL.ZS?](http://api.worldbank.org/countries/all/indicators/SP.RUR.TOTL.ZS?date=2000:2015&format=csv)

[date=2000:2015&format=csv](http://api.worldbank.org/countries/all/indicators/SP.RUR.TOTL.ZS?date=2000:2015&format=csv). It will immediately download the csv file for the rural population in all countries from 2000-2015.



1. Sara-Jayne Terp [↑](#footnote-ref-1)
2. XPATHS Resource: http://www.w3schools.com/xpath/xpath\_syntax.asp [↑](#footnote-ref-2)
3. Open API definition [↑](#footnote-ref-3)