Acquiring geographical data with web harvesting

K Dramowicz
Centre of Geographic Sciences, NSCC, Lawrencetown, Nova Scotia, Canada
Email: konrad.dramowicz@nscc.ca

Abstract. Many websites contain very attractive and up to date geographical information. This information can be extracted, stored, analyzed and mapped using web harvesting techniques. Poorly organized data from websites are transformed with web harvesting into a more structured format, which can be stored in a database and analyzed. Almost 25% of web traffic is related to web harvesting, mostly while using search engines. This paper presents how to harvest geographic information from web documents using the free tool called the Beautiful Soup, one of the most commonly used Python libraries for pulling data from HTML and XML files. It is a relatively easy task to process one static HTML table. The more challenging task is to extract and save information from tables located in multiple and poorly organized websites. Legal and ethical aspects of web harvesting are discussed as well. The paper demonstrates two case studies. The first one shows how to extract various types of information about the Good Country Index from the multiple web pages, load it into one attribute table and map the results. The second case study shows how script tools and GIS can be used to extract information from one hundred thirty six websites about Nova Scotia wines. In a little more than three minutes a database containing one hundred and six liquor stores selling these wines is created. Then the availability and spatial distribution of various types of wines (by grape types, by wineries, and by liquor stores) are mapped and analyzed.

1. Web harvesting: terminology and some ethical issues

Websites are the most important repositories of information that have ever existed. Information from websites is fundamental for all our activities, including communication, marketing, research, enjoyment, pleasure, etc. Vast and free information is scattered across countless web pages in various formats and often poorly defined structures. According to the online survey from February 29, 2016 [3], the total number of active websites was more than 996 million. Web harvesting, also called web data extraction or web scrapping, is a computer-based technique for extracting information from websites. In recent years, new automatic methods of browsing websites has grown very rapidly. The entire procedure for automatic browsing works in exactly the same way as browsing performed by human beings, including the simplest copy and paste method. However, the computer-based way of reading websites is much quicker than its human-based equivalent and also it can be very demanding for servers to process. Examples of web harvesting include online price comparison, weather data monitoring, website change detection, web mashup, web data integration, detecting changes in inventories, etc. With the help of GIS, web harvesting can be combined with web mapping and spatial analysis.

Web farming or web mining are less related to web harvesting. Web farming refers to setting up web servers, deploying content to them, and managing them. Whereas web harvesting focuses on extracting information from websites, not necessarily including analysis, web mining refers more to employing data mining techniques, emphasizing data analytics, and extracting knowledge from web data. Web mining techniques include ranking metrics (page quality and relevance), robot detection and filtering (human vs. nonhuman), analyzing so-called information scent, creating user profiles,
measuring interests, creating online bibliometrics, etc. Web harvesting very often represents the first step before starting web mining.

Web harvesting is just another way of gathering data and at the same time it may help a given website to reach more users. Web harvesting requires a lot of effort, skills, technology, and funds to be successful. Websites often have poorly structured HTML code, which makes the manual data extraction extremely difficult and time consuming. The role of web harvesting software is to process unstructured web data (usually in the HTML format) and store the well-structured information in a database or spreadsheet. There are different types and levels of web harvesting software tools. The simplest solution, offered by some companies, is to extract data online by supplying the URL of a website where there are tables full of data. This solution instantly turns a web page into data. Data extracted from a table can be saved, for example, in the CSV format. However, the most common solution is to install the web parser and perform web harvesting from the client side. There is an abundance of software solutions available for this purpose. Some of these solutions are rather expensive (more than $1,000 for one use of software), however, many of them are free. The most costly option is to outsource web harvesting to specialized companies, which offer post-processing of extracted data including, for instance, data aggregation, search engine rankings, social media and brand monitoring, financial and market research, etc.

Many websites contain abundant and valuable information that is difficult to harvest manually [2]. Therefore software applications (called bots) that run automated tasks over the internet are used. Bots, which systematically browse WWW for web indexing are called web crawlers. Web search engines use web crawlers to update the content and indexing of their websites. Web crawlers can copy all the visited pages for later processing using search engines and they are frequently used for web harvesting. Each server may have a file called robots.txt containing rules for crawling which the bot is supposed to obey. The website then uses the robots exclusion standard to communicate with web crawlers and other web robots using the robots.txt file. This standard informs the robot about which areas of the website should not be processed or even scanned. Anti-bot and anti-scraping technology is frequently used instead of applying legal barriers to stop web harvesting. With this technology, web servers may control how frequently a certain IP requests a website and then an access can be temporarily denied. Some websites require authentication before the user can access their content and if while entering a website there is no presence of real human, the logging process may stop. Many start-up companies proliferate on web harvesting from other businesses’ websites because it is a cheap and powerful way to gather data without the need for partnerships. These companies may have their own built or outsourced technology for acquiring a variety of data from websites. In contrary, big companies use web harvesting for their own advantages and at the same time they are against others who harvest their web pages. Generally, harvesting non-copyrighted or public domain data should not be classified as trespass to chattels, which is rather a fuzzy legal term and in various countries can be interpreted in different ways. The following are five basic web harvesting rules to obey:

1. Only public content can be harvested, in addition to obeying copyright policy.
2. Terms of use should be observed.
3. Data of a personal nature should not be used, otherwise it will lead to the invasion of privacy. The invasion of privacy occurs when information about individuals is harvested, used or disseminated without their consent or knowledge.
4. Data collected for a specific purpose cannot be used for another purpose (for example, for selling personal data as a commodity).
5. Amount of data extracted and frequency of harvesting should be reasonable.

2. Using the Beautiful Soup for harvesting the Good Country Index website

Beautiful Soup [1] is a Python library for pulling data out of HTML and XML files. It works with any parser and commonly saves programmers’ hours or days of work. There is also the equivalent in R
programming: the package rvest, inspired by Beautiful Soup, which helps to harvest web data with R by parsing tables, navigating around a website, and extracting all information from a web page.

The Good Country Index shows the cumulative contribution of countries towards the benefit of the world in the technological, cultural, peace-related, word orderliness, environmental, prosperous, and health wellbeing aspects (http://www.goodcountry.org/overall). The full information about the Good Country Index does not come from a single web page. Overall, there are eight Good Country web pages for overall index and seven corresponding categories and each of them should be opened and processed.

The procedure of harvesting, processing, and mapping the Good Country Index data consists of many steps, including web harvesting with Python’s module Beautiful Soup, processing data in Python, geoprocessing and mapping in ArcGIS with the Python module arcpy. Similar procedures were used for analyzing Nova Scotia wines and these procedures are explained with more details. The final map, created in a matter of seconds, shows countries ranked by category Culture using online data from the Good Country Index website (figure 1). The batch mode processing allows the user to open eight Good Country Index web pages and load their data into a GIS database in less than one minute.

![Figure 1: Countries ranked by the Good Country Index, category Culture](image)

3. Using the Beautiful Soup for harvesting websites about wines from Nova Scotia

In 2014, there were ten wineries belonging to the Winery Association of Nova Scotia producing 136 different wines from 37 varieties of wine grapes and other ingredients (fruits, maple syrup). These wines are available across the province of Nova Scotia in 106 Nova Scotia Liquor Corporation (NSLC) stores. Information about how many bottles of a given wine are available in which NSLC store was harvested from 136 web pages, each corresponding to one wine code. These web pages are dynamic, showing NSLC wine inventory updated in real time. The following is the list of major steps performed (figure 2):

1. Data preparation:
   a. Obtain codes for each Nova Scotia wine product offered in liquor stores
   b. Create the list of wine grapes and associate wine grapes with wine names
c. Obtain addresses of liquor stores and geocode them
d. Obtain URL for each winery’s website
2. In ArcGIS, create a script tool with such parameters as: user’s location (see step 8), Boolean options for selecting one wine grape from the list, for selecting one winery from the list of wineries where a given wine is produced (see step 10), and for opening a web page showing the list of liquor stores and the number of bottles of a selected wine (see step 5)
3. With Python, create two parallel lists: wine names and corresponding product codes, pass the list of wine names to Tkinter (Python GUI package)
4. Use Tkinter for displaying the list of wines and getting from the user the name of one selected wine
5. Append the wine code to the end of the URL address of the Nova Scotia Liquor Corporation and, optionally, open the corresponding web page using the imported Python module webbrowser
6. Use the Python and its module Beautiful Soup for:
   a. Accessing a web page for reading its contents with the imported Python module urllib2
   b. Browse all rows in every table in the processed web page and count the number of corresponding columns in each table for finding the table with exactly seven columns
   c. Unpack information from a given row into the variables, including store ID and the number of bottles
   d. Convert the Unicode values, representing store ID and the number of bottles, to the string; convert the number of bottles from the string to integer
   e. Accumulate the number of bottles from all rows (stores)
   f. Create two new parallel lists with the store IDs and number of bottles
7. In ArcGIS, use the imported Python module arcpy and:
   a. Apply an SQL statement for extracting liquor stores where a given wine is available
   b. Create an empty DBF table with two fields: store ID and number of bottles
   c. Use the insert cursor for looping and adding one row at a time, then populate empty fields with values from two existing parallel lists
   d. Join the created DBF table with a shapefile containing stores
   e. Display the total number of bottles of a given wine in all liquor stores
   f. Display a map showing stores, where proportional circles represent the number of bottles
   g. Use the arcpy submodule mapping for removing previous layers from the map
8. Optionally, use the Network Analyst extension to ArcGIS for performing the following tasks:
   a. Use the road network, assign speed values to road types, calculate length and the corresponding driving time for each road segment
   b. Create the road network dataset, define fields for time and length impedances, and a field with street names for driving directions
   c. Use the Closest Facility solver for finding the closest store where a given wine is available
   d. Load incident (user’s location)
   e. Load facilities (selected liquor stores)
   f. Solve the problem, extract the distance and driving time, convert distance to meters and time to minutes and seconds
   g. Map the shortest route
   h. Show driving directions
9. Use Tkinter again for displaying the output form with the wine name, number of bottles, store address, distance, and driving time
10. Optionally, if the user wants to begin with selecting wine grapes, the following additional tasks will be completed using Python and Tkinter:
a. Display the list of wine grapes for selecting one name
b. Find the corresponding list of wines produced from selected grapes and display such a list for selecting one wine
c. Find the corresponding wineries where a given wine is produced and display the list for selecting one winery
d. Open the selected winery’s web page

Figure 2: Process diagram showing used technology and corresponding step numbers

The user can select wine name from the list of wines and select interactively user’s location on a map. As a result, a given wine is added to a mappable geocoded database with liquor stores and the shortest path from the user’s location to the closest store and driving directions are shown (figure 3).

Figure 3: The output from the geoprocessing script tool for the selected wine Avondale Sky Bin 33
The whole process of opening 136 web pages, each with information about the number of bottles of a particular Nova Scotia wine in each of 106 liquor stores and loading this information to one database table with 106 rows and 136 columns took only a little longer than three minutes. The resulting table, after being joined to another table with 37 grapes used for producing wines, allows the user to select one grape variety, then one wine from the list of wines made from the already selected grapes and then find a winery producing a given wine. Figure 4 shows the interactive menu used to select grapes, wines, and wineries. As the result, the number of bottles of a given wine can be grouped by a winery or by liquor store. Then some analyses comparing the availability of various types of wines can also be conducted.

![Figure 4: Selecting grapes, wines, and wineries](image)

The number of bottles of wines in liquor stores is proportional to the corresponding market size (figure 5). However, based on figure 6, the proximity of a winery to a liquor store is not observed. In larger markets (Halifax/Dartmouth, Sydney, Wolfville/New Minas/Kentville) wines from many wineries are available. In other stores the prevailing producer is the Jost Vineyards and its sister company Gaspereau Vineyards.

![Wineries: Number of Bottles in NSLC Stores by Type of Wine](image)

![NSLC Stores: Total Number of Bottles of NS Wines](image)

*Figure 5: Number of bottles of Nova Scotia wines in liquor stores by Nova Scotia winery and by store*
Figure 6: Number of bottles in Nova Scotia wines in liquor stores by Nova Scotia winery by stores

Figure 7 presents four maps with the number of bottles of white, red, rose, and sparkling wines produced by Nova Scotia wineries and available in liquor stores across the province [4]. The proportional circles show the number of bottles of wines produced in Nova Scotia and available in each liquor store. Figures 6 and 7 show that in every Nova Scotia liquor store there were more bottles of white than red wines. Rose and sparkling wines were less popular. The Jost Vinyards from Malagash is the largest wine producer in Nova Scotia.

Figure 7: Number of bottles of Nova Scotia white, red, rose, and sparkling wines in liquor stores
4. Final remarks
The presented examples of extracting geographic information from websites confirm the usefulness of web harvesting combined with Geographic information Systems (GIS). Geographic information can be in the form of coordinates or addresses (like with liquor stores or wineries), which can be geocoded and mapped. Also, this information may contain just names of geographic features (like country names from the Good Country Index websites). This extracted information, placed into a database table, can be related (joined) to existing maps of objects (countries). Web harvesting provides invaluable tool for acquiring practically any geographic information from online sources. This technology can be embedded inside GIS tools allowing the user to decide which portion of information should be extracted for processing. After harvesting multiple pages, all acquired information can be linked together into one database, if they have something in common. Then GIS can be used for mapping, spatial analysis, location analysis, spatially-related data analytics, etc. Also, repeated web harvesting may provide information for analyzing changes. For example, information about the number of bottles of Nova Scotia wines in liquor stores acquired on April 5, 2014 can be compared with another day, since the inventory of bottles changes every minute and all these dynamic changes are reported online.

The role of geographic information available online will be even more critical in the future. Easy access to GPS, location-based services, wireless communication, and sensors will establish a new role of GIS for ordinary citizens. Unfortunately, usually geographic information accessible online is not user-friendly for extracting. Moreover, for most GIS users, the process of web harvesting is too time consuming and difficult to be conducted at all. This project demonstrates the prototype solution when web harvesting is combined with GIS. Web harvesting is able to gather online information and process it. However, analyzing processed data is beyond the scope of web harvesting. On the other side, obtaining online data does not represent a typical GIS functionality. GIS is capable of processing, mapping, and performing various types of geographical and statistical analysis of georeferenced data. GIS can benefit from integrating new technology on its data acquisition and data analysis both stages. Web harvesting proves to be an important contributor to GIS data acquisition functionality. On the other side, data mining can enhance analytical capabilities of GIS. Further integration of web harvesting with GIS is recommended as a synergic solution.

5. References
[1] Beautiful Soup 4.4.0 Documentation. http://www.crummy.com/software/BeautifulSoup/bs4/doc
[2] Hartley B 2014. The Ultimate Guide to Web Scrapping. https://blog.hartleybrody.com/guide-to-web-scrapping/
[3] Internet Live Stats. http://www.internetlivestats.com/total-number-of-websites
[4] Slaunwhite L 2014 Mapping and Querying Wine Products in Nova Scotia. Major Research Project in GIS for Business (supervised by K. Dramowicz), Centre of Geographic Sciences, NSCC, Lawrencetown, Nova Scotia, Canada. This is the source of maps 5-7 presented in this article.