Global Investigative Journalism Network (https://gijn.org/2015/08/11/web-scraping-a-journalists-guide/)

Web Scraping: A Journalist's Guide

By: **NAEL SHIAB** | August 11, 2015

(http://gijn.wpengine.netdna-cdn.com/files/2015/07/Backlit_keyboard-1024x576.jpg) Do you remember when Twitter lost $8 billion in just a few hours (http://www.bbc.com/news/technology-32511932) earlier this year? It was because of a web scraper, a tool companies use—as do many data reporters.

A web scraper is simply a computer program that reads the HTML code from webpages, and analyze it. With such a program, or "bot," it's possible to extract data and information from websites.

Let's go back in time. Last April, Twitter was supposed to announce its trimestrial financial results once the stock markets closed. Because the results were a little bit disappointing, Twitter wanted to avoid a brutal confidence loss from the traders. Unfortunately, because of a mistake, the results were published online for 45 seconds, when the stock markets were still open.

These 45 seconds allowed a bot programmed to web scrape to find the results, format them and automatically publish them on Twitter itself. (Nowadays, even bots have scoops from time to time!)

Selerity
@Selerity

Follow

#BREAKING: Twitter $TWTR Q1 Revenue misses estimates, $436M vs. $456.52M expected

4:07 PM - 28 Apr 2015

157 88

Once the tweet was published, traders went crazy. It was a disaster for Twitter. The bot's company, Selerity (http://www.seleritycorp.com/), specializes in real-time analysis, and became the target of many critics. The company explained the situation a few minutes later.

Selerity
@Selerity

Follow

Today's $TWTR earnings release was sourced from Twitter's Investor Relations website investor.twitterinc.com. No leak. No hack.

4:34 PM - 28 Apr 2015

570 312

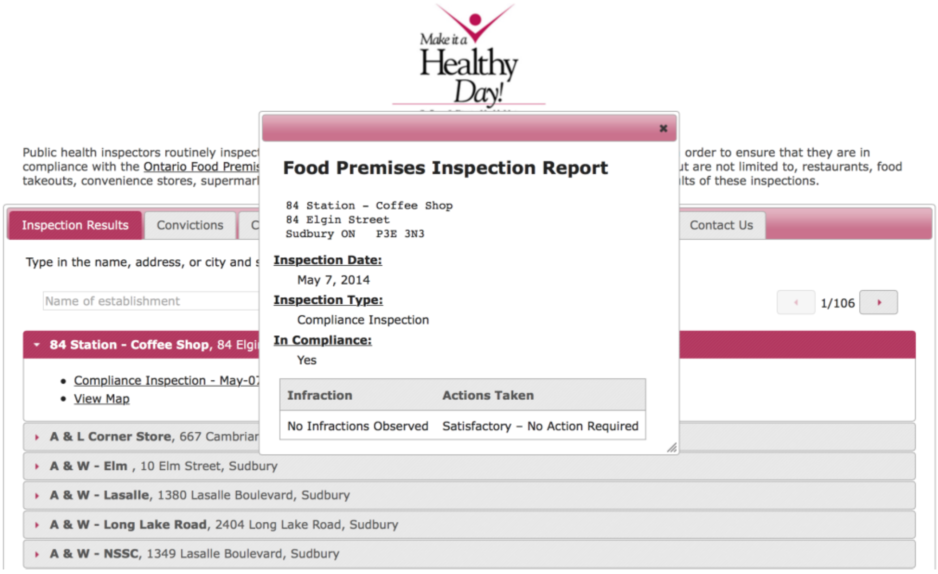
For a bot, 45 seconds is an eternity. According to the company, it took only three seconds for its bot to publish the financial results (http://arstechnica.com/business/2015/05/how-selerity-reported-twitters-2q15-earnings-before-twitter-did/) .

Web Scraping and Journalism

As more and more public institutions publish data on websites, web scraping has become an increasingly useful tool for reporters who know how to code.

For example: for a story for Journal Métro (http://journalmetro.com/actualites/national/789697/saq-des-centaines-de-produits-moins-chers-en-ontario/) , I used a web scraper to compare the price of 12,000 products from the Société des alcools du Québec with the price of 10,000 products of the LCBO in Ontario.

Another time, when I was in Sudbury, I decided to investigate food inspections in restaurants. All the results from such investigations are published on the Sudbury Health Unit (http://inspectionresults.sdhu.com/) 's website. However, it's impossible to download all the results; you can only verify the restaurants one by one.



Public health inspectors routinely inspect ... order to ensure that they are in compliance with the Ontario Food Premis... ...ut are not limited to, restaurants, food takeouts, convenience stores, supermar... ...lts of these inspections.

Type in the name, address, or city and s

Name of establishment

Food Premises Inspection Report

```
84 Station - Coffee Shop
84 Elgin Street
Sudbury ON   P3E 3N3
```

Inspection Date:
May 7, 2014

Inspection Type:
Compliance Inspection

In Compliance:
Yes

Infraction	Actions Taken
No Infractions Observed	Satisfactory – No Action Required

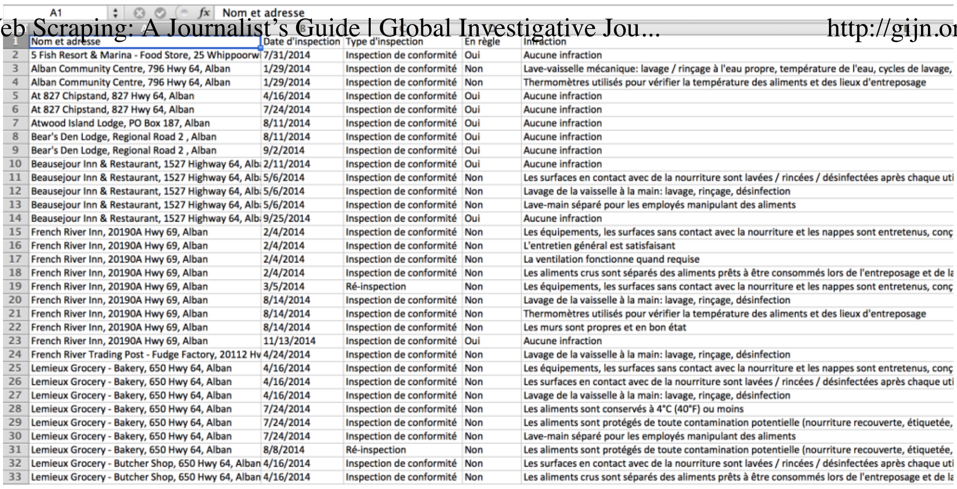
- **84 Station - Coffee Shop**, 84 Elgi
  - Compliance Inspection - May-07
  - View Map
- **A & L Corner Store**, 667 Cambrian
- **A & W - Elm** , 10 Elm Street, Sudbury
- **A & W - Lasalle**, 1380 Lasalle Boulevard, Sudbury
- **A & W - Long Lake Road**, 2404 Long Lake Road, Sudbury
- **A & W - NSSC**, 1349 Lasalle Boulevard, Sudbury

I asked for the entire database where the results are stored. After a first refusal, I filed a freedom-of-information request—after which the Health Unit asked for a $2,000 fee to process my request.

Instead of paying, I decided to code my own bot, one that would extract all the results directly from the website. Here is the result:

Coded in Python, my bot takes control of Google Chrome with the Selenium (http://selenium-python.readthedocs.org/) library. It clicks on each result for the 1600 facilities inspected by the Health Unit, extracts the data and then sends the information into an Excel file.

To do all of that by yourself would take you weeks. For my bot, it was one night of work.


A1 fx Nom et adresse

1	Nom et adresse	Date d'inspection	Type d'inspection	En règle	Infraction
2	5 Fish Resort & Marina - Food Store, 25 Whippoorwi	7/31/2014	Inspection de conformité	Oui	Aucune infraction
3	Alban Community Centre, 796 Hwy 64, Alban	1/29/2014	Inspection de conformité	Non	Lave-vaisselle mécanique: lavage / rinçage à l'eau propre, température de l'eau, cycles de lavage,
4	Alban Community Centre, 796 Hwy 64, Alban	1/29/2014	Inspection de conformité	Non	Thermomètres utilisés pour vérifier la température des aliments et des lieux d'entreposage
5	At 827 Chipstand, 827 Hwy 64, Alban	4/16/2014	Inspection de conformité	Oui	Aucune infraction
6	At 827 Chipstand, 827 Hwy 64, Alban	7/24/2014	Inspection de conformité	Oui	Aucune infraction
7	Atwood Island Lodge, PO Box 187, Alban	8/11/2014	Inspection de conformité	Oui	Aucune infraction
8	Bear's Den Lodge, Regional Road 2 , Alban	8/11/2014	Inspection de conformité	Oui	Aucune infraction
9	Bear's Den Lodge, Regional Road 2 , Alban	9/2/2014	Inspection de conformité	Oui	Aucune infraction
10	Beausejour Inn & Restaurant, 1527 Highway 64, Alb	2/11/2014	Inspection de conformité	Oui	Aucune infraction
11	Beausejour Inn & Restaurant, 1527 Highway 64, Alb	5/6/2014	Inspection de conformité	Non	Les surfaces en contact avec de la nourriture sont lavées / rincées / désinfectées après chaque uti
12	Beausejour Inn & Restaurant, 1527 Highway 64, Alb	5/6/2014	Inspection de conformité	Non	Lavage de la vaisselle à la main: lavage, rinçage, désinfection
13	Beausejour Inn & Restaurant, 1527 Highway 64, Alb	5/6/2014	Inspection de conformité	Non	Lave-main séparé pour les employés manipulant des aliments
14	Beausejour Inn & Restaurant, 1527 Highway 64, Alb	9/25/2014	Inspection de conformité	Oui	Aucune infraction
15	French River Inn, 20190A Hwy 69, Alban	2/4/2014	Inspection de conformité	Non	Les équipements, les surfaces sans contact avec la nourriture et les nappes sont entretenus, conç
16	French River Inn, 20190A Hwy 69, Alban	2/4/2014	Inspection de conformité	Non	L'entretien général est satisfaisant
17	French River Inn, 20190A Hwy 69, Alban	2/4/2014	Inspection de conformité	Non	La ventilation fonctionne quand requise
18	French River Inn, 20190A Hwy 69, Alban	2/4/2014	Inspection de conformité	Non	Les aliments crus sont séparés des aliments prêts à être consommés lors de l'entreposage et de la
19	French River Inn, 20190A Hwy 69, Alban	3/5/2014	Ré-inspection	Non	Les équipements, les surfaces sans contact avec la nourriture et les nappes sont entretenus, conç
20	French River Inn, 20190A Hwy 69, Alban	8/14/2014	Inspection de conformité	Non	Lavage de la vaisselle à la main: lavage, rinçage, désinfection
21	French River Inn, 20190A Hwy 69, Alban	8/14/2014	Inspection de conformité	Non	Thermomètres utilisés pour vérifier la température des aliments et des lieux d'entreposage
22	French River Inn, 20190A Hwy 69, Alban	8/14/2014	Inspection de conformité	Non	Les murs sont propres et en bon état
23	French River Inn, 20190A Hwy 69, Alban	11/13/2014	Inspection de conformité	Oui	Aucune infraction
24	French River Trading Post - Fudge Factory, 20112 Hv	4/24/2014	Inspection de conformité	Non	Lavage de la vaisselle à la main: lavage, rinçage, désinfection
25	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	4/16/2014	Inspection de conformité	Non	Les équipements, les surfaces sans contact avec la nourriture et les nappes sont entretenus, conç
26	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	4/16/2014	Inspection de conformité	Non	Les surfaces en contact avec de la nourriture sont lavées / rincées / désinfectées après chaque uti
27	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	4/16/2014	Inspection de conformité	Non	Lavage de la vaisselle à la main: lavage, rinçage, désinfection
28	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	7/24/2014	Inspection de conformité	Non	Les aliments sont conservés à 4°C (40°F) ou moins
29	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	7/24/2014	Inspection de conformité	Non	Les aliments sont protégés de toute contamination potentielle (nourriture recouverte, étiquetée,
30	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	7/24/2014	Inspection de conformité	Non	Lave-main séparé pour les employés manipulant des aliments
31	Lemieux Grocery - Bakery, 650 Hwy 64, Alban	8/8/2014	Ré-inspection	Non	Les aliments sont protégés de toute contamination potentielle (nourriture recouverte, étiquetée,
32	Lemieux Grocery - Butcher Shop, 650 Hwy 64, Alban	4/16/2014	Inspection de conformité	Non	Les surfaces en contact avec de la nourriture sont lavées / rincées / désinfectées après chaque uti
33	Lemieux Grocery - Butcher Shop, 650 Hwy 64, Alban	4/16/2014	Inspection de conformité	Non	Les aliments crus sont séparés des aliments prêts à être consommés lors de l'entreposage et de la

But while my bot was tirelessly extracting thousands of lines of code, one thought kept bothering me: what are the ethical rules of web scraping?

Do we have the right to extract any information found on the web? Where is the line between scraping, and hacking? And how can you ensure that the process is transparent for both the institutions targeted and the public reading the story?

As reporters, we have to respect the highest ethical standards. Otherwise, how can readers trust the facts we report to them?

Unfortunately, the code of conduct of the Fédération professionnelle des journalistes du Québec (http://www.fpjq.org/deontologie/guide-de-deontologie/#pt5) , adopted in 1996 and amended in 2010, is getting old and brings no clear answers to all my questions.

The ethics guidelines (http://www.caj.ca/ethics-guidelines/) of the Canadian Association of Journalists, although more recent, doesn’t shed much light on the matter, either.

As Université de Québec à Montréal journalism professor Jean-Hugues Roy (https://twitter.com/jeanhuguesroy) says it: “These are new territories. There are new tools that push us to rethink what ethics are, and the ethics have to evolve with them.”

So, I decided to find the answers by myself, by contacting several data reporters in the country.

Stay tuned; the results from that survey will be published in a following instalment.

Note: If you’d like to try a web scrape yourself, I published a short tutorial last February (http://naelshiab.com/members-parliament-web-scraping/) . You will learn how to extract data from the Parliament of Canada website!

This post originally appeared on J-Source.CA (http://j-source.ca/article/journalists-guide-web-scraping) and is reprinted with permission. To see this story in Chinese, check GIJN's Chinese-language site. (http://cn.gijn.org/2015/08/21/%E6%96%B0%E9%97%BB%E4%BA%BA%E7%BD%91%E7%BB%9C%E6%95%B0%E6%8D%AE%E9%87%87%E9%9B%86%E5%85%A5%E9%97%A8/)

(http://gijn.wpengine.netdna-cdn.com/files/2015/07/nael.jpg) Nael Shiab is an MA graduate of the University of King's College digital journalism program. He has worked as a video reporter for Radio-Canada and is currently a data reporter for Transcontinental. @NaelShiab (https://twitter.com/NaelShiab)