Web Scraping and BLS

Federal Committee on Statistical Methodology
Research Conference March 7, 2018

Bill Thompson
Overview

- Background
- Experiences at BLS
- Challenges
- Next Steps
Data Collection Challenges

■ Reluctant respondent
  o Burden
  o Trust

■ Data Collection
  o Costly to support entire infrastructure
Data Collection Challenges

- Called to automate our data collection
How BLS has Addressed Data Collection Challenges

- Purchased data sets
- Other Federal surveys
- Directly from corporation
- Scraping websites
Data Collection Opportunity

- Information on the internet

- Anything, well almost anything, we collect manually could be collected automatically
Web Scraping

Web scraping is the process of extracting data from websites, typically through the use of a software program that simulates human exploration of a website(s).
Getting Web Data

- Web scraping etc...
  - Extract data (prices) from web page (HTML)
    - Human readable page -> machine readable data
  - Using APIs (Application Programming Interfaces)
    - Code (provided by source) get machine readable data directly

- Next step--data-traffic efficient applications
Web Scrape Benefits

- Reduces collection costs
- Reduces respondent burden
- Improves timeliness
- Increases sample size
- Increases data quality
- Allows for evaluation & improvement
Web Scraping

- In the beginning ...

  - Obtain price and characteristic data for hedonic modeling

  - Evaluate as a replacement for data collection
Web Scraping

Then there were multiple efforts...

- Retailer web sites, web scraping & API
- Price aggregators
- Obtain data from other statistical agencies
- Location services to refine address information
- Job postings & The Conference Board
Web Scraping Successes

- Fatal Work Injuries & Google Alerts
- Office of Productivity & Technology & Federal Agencies
- Prices
Census of Fatal Occupational Injuries (CFOI)

- Fatal Work Injuries & Google Alerts
  - Google Alerts
  - Developed CFOI Public Data Management System (C-PDMS)
  - C-PDMS is being used by for CFOI production purposes.
Office Of Productivity & Technology (OPT)

Web Scrape Throughout OPT

- Multifactor Productivity
  - 40% of the input data in non-manufacturing & manufacturing
  - Bureau of Economic Analysis (BEA)

- Industry Productivity
  - United States Geological Survey (USGS)
  - & Center Disease Control (CDC)
  - & Bureau of Labor Statistics (BLS)

- Reaching out to BEA for more source data into the API
Price research

- A fairly significant amount of research into web scraping activity in potential support of the Consumer Price Index

- Overall objective of automating collection of prices and product information to create price indexes
Consumer Price Index & Division of Price and Index Number Research

- Current Price Research
  - Using the API (with permission)
    - From retailers to collect prices and characteristic information on a weekly basis
    - From online price aggregator
  - Using DPINR’s automated data collection
    - From retailers (with permission)
What have we learned?

BLS can systematically harness information on the internet for statistical use...clearly BLS has transitioned out of the ‘proof of concept’ stage.
Challenges

- Infrastructure
- Appropriate skills
- Research
- And...
Policy Issues

- Information on website=publicly available data?
- How does confidentiality apply?
- Optics & future respondent cooperation matter
In the Meantime

■ Except for a few situations
  ○ CPI has scaled back effort
  ○ Secure permission prior to web scrape

■ Guidance
  ○ Developing a business plan for each planned use
  ○ Developing a guidance/decision

■ Discussion about current policy
Next Steps

- Policy concerns
- Begin to consider a centralized approach
- Continue to seek out opportunities
Contact Information

Bill Thompson
Producer Price Index
thompson.bill@bls.gov

David Friedman
Associate Commissioner Office of Price & Living Conditions
friedman.david@bls.gov