Web scraping meets survey design: combining forces

Statistics Netherlands

Olav ten Bosch

BigSurv18 conference, Barcelona, 26-10-2018

Contents

- Scraping, why, what?
- Survey data vs.
  web scraped data
- Technical & legal
- Webscraping phases
- Combining forces
- Wrap up

Scraping projects: why?

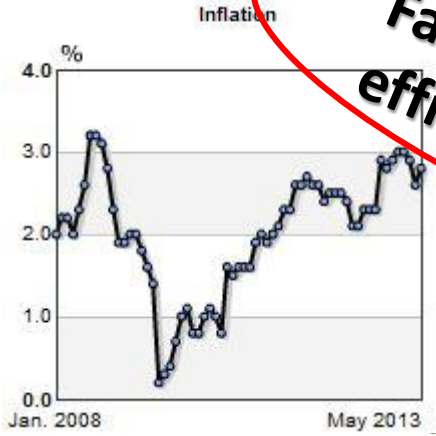
Administrative sources

- Tax, social security
- Municipalities/ Provinces
- Supermarkets
- …

Internet sources

- …
- Surveys

Less!!!



Faster,better,more efficient

New

14-06-2013	Exports shrink
14-06-2013	Calendar
13-06-2013	Retail turnover 0.6 percent lower
13-06-2013	Large drop in turnover for car and motorcycle trade

New indicators

Scraping projects: what?

Internet sources

Some use cases

- Internet prices for CPI, clothing, airline tickets, restaurants
- Real estate sites for housing statistics
- Job portals for job market statistics
- Consumer sites for second-hand goods for early economic indicators
- Wikipedia for improving the business register
- Enterprise websites for Ecommerce, social media, NACE, etc.

Survey data

Web scraped data

Processing of survey data

Processing of web scraped data

Survey data vs Scraped data

- Designed by NSI
- Well structured
- Relatively stable
- High quality
- Small volumes, processed in waves
- Statistical classifications

- Not designed by NSI
- Sometimes messy
- May change any time
- Quality depends on source
- Can be big, continuous processing & monitoring
- Concepts used in practice

Technical & Legal

Technical: static/dynamic, HTML/API, technology changes, open source

⇒ *If you can see it, we can scrape it*
(and sometimes even more)

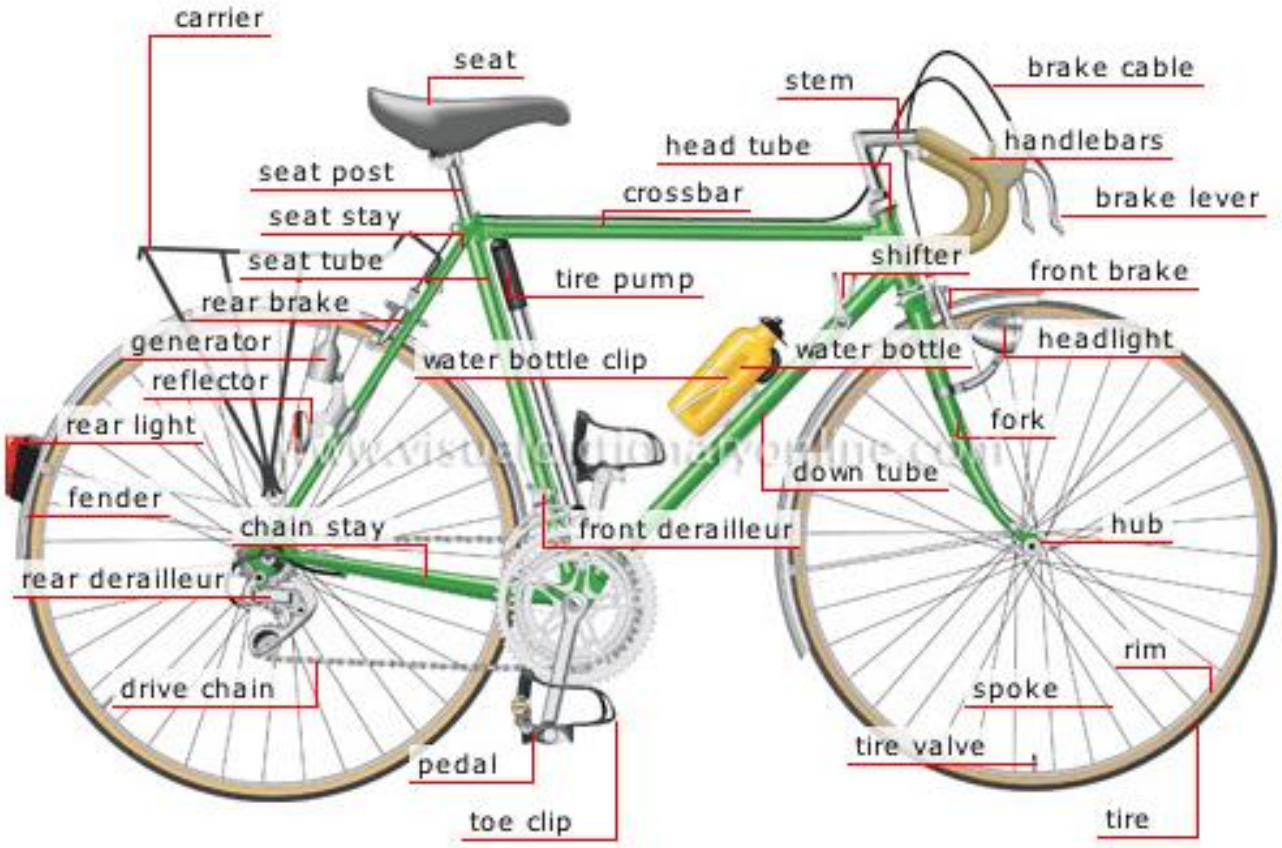
Legal: statistical law, intellectual property rights, privacy, netetiquette:

=> *Crucial but manageable for an NSI*

Three phases in web scraping

1. ***Site analysis phase***:
   - Examine the web source(s)
   - Programmability, volume, volatility, legal, originality, uniqueness, detail, navigation
2. ***Data analysis and design phase***:
   - Set up a test data stream
   - Stability, redundancy, plausibility, itentifiability, combinability, role in statistical process
3. ***Production phase***:
   - Data used in production
   - Monitoring, organisation, maintenance, communication with site owners

Site analysis phase



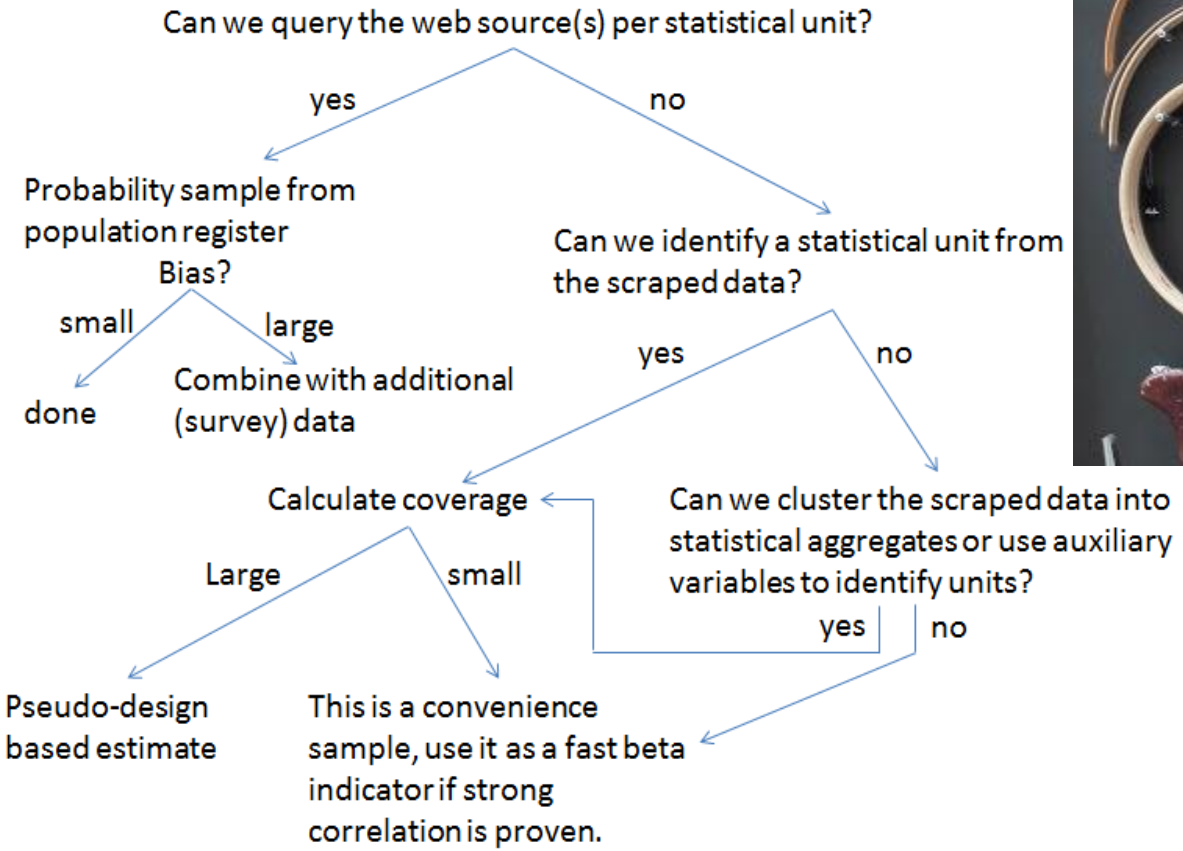
Data analysis and design phase

Production phase

Combing forces: surveys and scraping

- Can we use the best of both worlds?
- Questions in site analysis and design phase:
  - Can we query the web source(s) per statistical unit from our register?
  - Can we identify a statistical unit from the scraped data directly or indirectly?
  - Can we cluster scraped data into aggregates that link to known statistical units?
  - Can we use the data for calculating a fast beta indicator?

Generic workflow, a first try



Wrap up

- Web scraped data differs from survey data in many aspects
- Both have pros and cons
- Webscraping: technical and legal aspects are crucial but manageable
- We see three phases in setting up webscraping
- A first general workflow on combining survey and web data has been sketched

Thank you, questions, ideas, suggestions

?

Olav ten Bosch
obos@cbs.nl

Something you may also be interested in:

Curated list of software for official statistics

awesome

www.awesomeofficialstatistics.org

Pictures from author and (cc or with permission): pxhere.com, inframarks.nl, 247fietsen.nl