Webscraping in FinEco Research: Risks and Opportunities

Cynthia A. Huang

Econometrics & Business Statistics, Monash University

Introduction

About Me

- 💆 PhD Candidate in Monash EBS, affiliated with
 - NUMBATS (EBS), SoDa Labs (ECON), EmVis (Faculty of IT)
 - Monash Data Futures Institute
- Researching principles and methods for "alternative" data
 - conceptual and practical data provenance tools for harmonised multi-source datasets
 - adapting web-scraped retail product & price data for public health research
 - Statistical properties of alternative data, grammar of graphics
- Open and reproducible research tools, research software design
 - Quarto, git, replication packages
 - NUMBATS Hacky hour, community building

Objectives

Research opportunities & risks

- Big data, novel data, alternative data
- Operational vs. scientific risk
- Web technologies and webscraping methods
- Ethical & legal risk

How to code a web-scraper
How to code big-data analysis
Project level considerations
Resourcing web-scraping projects (skill sets, RA hires etc.)

Web scraping—verb.

Cambridge dictionary:

the activity of **taking** information from a <u>website or computer screen</u> and **putting** it into an <u>ordered document</u> on a computer



Webscraping in FinEco Research: Risks and Opportunities, Cynthia A. Huang, 5 Apr 2024

Web scraped data—noun.

Cambridge dictionary:

the activity of taking **information from a website or computer screen** and putting it into an ordered document on a computer



Common properties of web scraped data



- Some degree of messiness
 - Collection, preparation
 - Analysis, statistical properties
- Non-trivial parsing from raw web extraction to analysisready data
- Murky data dimensions/quality
 - How many records?
 - How many observational units?
 - Complete/missing observations?

Research Opportunities & Risks

Big, novel, alternative (data) opportunities

- The internet is a rapidly expanding universe of (collectable*) data
 - Multiple modalities: text, images, video...
 - Multi-source harmonization: shopping aggregators, knowledge graphs
 - Novel signal sources: social media sentiment, digitial economy indicators
- Lower* cost than traditionally available data collection methods
 - Digital vs. Analogue access and delivery
 - Automated data extraction vs. Manual data entry
- Volume, Velocity, Variety... Veracity?
 - Noise vs. signal for data mining, prediction, casual inference
 - Collecting data from a Complex System vs. Library

Research Risk: Traditional Data

Operational Risk

- Availability and Access
- Cost
- Legal and ethical

Scientific Risk

- Suitable analysis methods
- Signal-to-noise ratio (in context)
- Data quality

Research Risk: Web Scraped Data

Operational Risk

- Cost = Time
- Availability and access
 - Sampling bias
 - Unstable web interfaces
- Legal and ethical
 - Usage terms

Scientific Risk

- Data quality
 - Completeness
 - Accuracy

Signal-to-noise ratio

- Big Data Paradox (Bradley et al., 2021)
- Suitable analysis methods

Data Collection Scope

Conceptual Scope



Resource Sink

Example dataset description:

Daily stock prices

- Conceptually and operationally open-ended
- Unclear dataset dimensions
 - Which firms?
 - What time period?



Maintenance Trap

Example dataset description:

Daily prices for alcoholic beverages sold online by retailers with physical stores in Australia

- Conceptually limited, but operationally open-ended
- implies *ongoing* collection:
 - Repair "broken" scrapers
 - Upgrade storage & compute
 - (Cavallo & Rigobon, 2016)



Sampling bias

Example dataset description:

Daily prices for alcoholic beverages available online between Jan-Dec 2023

- Operationally limited, but conceptually open-ended
- What population is being sampled from?
 - Index coverage (Foerderer, 2023)



Research Quality Dataset(s)

- Conceptually and operationally limited
- Quality is *context* dependent!
- Requires deliberate design decisions
 - Filtering criteria (e.g. top products, largest retailers)
 - Conversion between observation and analysis resolution (e.g. from retailer to brand)

Business vs. Research Needs



Webscraping in FinEco Research: Risks and Opportunities, Cynthia A. Huang, 5 Apr 2024

Case Study

Adapting web scraped retail price data for public health research

Outsourcing web scraping



Data refinements & augmentation

Anomalies in cross-retailer harmonisation



Augmenting product attributes

- Product name
- Brand name
- Manufacturer
- Product category
- Volume per unit
- Alcohol by Volume
- Standard drinks per unit

Web technologies, access interfaces, and collection methods

Web scraping: Expectations vs. Reality



Web scraping: Expectations vs. Reality



Webscraping in FinEco Research: Risks and Opportunities, Cynthia A. Huang, 5 Apr 2024

Live Demo

- HTML elements
- API responses

•••	💅 Yahoo Finance - Stock Market L 🗙 🛛 🖉 https://www.danmurphys.com.a 🗙 🕂					~
\leftarrow \rightarrow C \textcircled{a}	○ A https://finance.yahoo.com	E - 公	${igodot}$	<u>ナ</u> ② む	289 🐣	
yahoo!	Search for news, symbols or companies	Q	C	Sign in		Mail
Watchlists My	y Portfolio Markets News Videos Screeners Personal Finance Cry	ypto Sectors Contact Us				
acti	vity shows market is swamping Trump's	for a mortgag a.Fw(b).Ell.D(b).C	(\$linkColor).Po	os(r).Z(2) 65 × 17.5	Change	% Change
Yahoo	e Finance		DJT Trump Media	48.66 & Technology Group	-13.30 Corp.	-21.47%
			UNH UnitedHealt	489.70 Group Incorporated	-5.00	-1.01%
	Analyst revamps Micron price target on memory chi	p demand 🖞	CXAI CXApp Inc.	6.15	+3.70	+151.02%
	outlook Micron's place in a key Al-adjacent market has analysts looking for a bi	ig revenue boost.	GOOG Alphabet Inc	156.50	+4.24	+2.78%
https://finance.yahoo.com/quo	ote/DJT Don't Count Samsung Out in SK Hynix Tops	s \$100 Billion in	MU	124.30	+6.41	+5.44%
🕞 🗘 Inspector 🕞 Con	isole D Debugger 🔨 Network {} Style Editor 🕥 Performance 🕼 Memory 🗄 Storage 📑	Accessibility 🎆 Application			99+ 🗖	0] ••• ×
Q Search HTML		+ 🗡	Rules	Layout Computed	Changes	Compatibil
	<pre></pre>	="Trump Media & Technology	 Filter Styles Pseudo-elem 	: nents	nov .cls +	* 🛈 🖹
	<pre>▶ Technology Group Corp."></pre>	Ell::after, [class*="Lind content: " font-size: height: 0; width: 0;	ato eClamp"]::after ::: ."; 0;	mic-light.4 {	41b9ceb9.css:1	
< 00%).Pos(r).Tbl(f).Bdcl(c).Bd	<pre></pre>	b).Ell.D(b).C(\$linkColor).Pos(r).Z(>	atomic .V\(h Ell::after, [class*="Lind visibility	\), <u>ato</u> eClamp"]::after <u>:</u> ; : hidden;	mic-light.4 {	41b9ceb9.css:1

Web technologies

- Webpages and websites
 - Styling: CSS
 - Static content: HTML/XML
 - Dynamic content: Javascript, API queries
- Web Protocols (HTTP/HTTPS)
- Application Programming Interface (APIs)
 - Query parameters
 - Response formats: JSON, CSV...



Web scraping methods: Indexing & Crawling

Figure 1: Web Scraping Process: From the Target Population to the Sample.

Web scraping entails two steps, *indexing* and *fetching*. In indexing, the target population is systematically registered. Indexing yields the frame in terms of a register of all units in the population, together with the URLs pointing to each unit. Fetching automatically visits each URL listed in the frame and downloads the resource at which it points, typically an HTML document.



Webscraping in FinEco Research: Risks and Opportunities, Cynthia A. Huang, 5 Apr 2024

Web scraping methods: Retrieval

Manual Copy-Paste

- Concurrent* collection & parsing
- Human errors and judgement

Interactive collection

- Automating mouse/keyboard inputs
- Via headless browsers
- Selenium, chromium etc.

Request-based extraction

- Request resources using machine readable methods
- HTTP CONNECT/GET
- API Queries

Web scraping methods: Parsing



An overview of HTTP - HTTP | MDN. (2023, December 16). https://developer.mozilla.org/en-US/docs/Web/HTTP/Overview

- HTML element extraction:
 - Tables, if you're lucky
 - Often requires detangling style information from metadata and data

• API response parsing:

- A way for computers to talk to computers without human readability/presentation layer
- Deciphering JSON keys can be tricky without access to dictionary

Collection Risk by Method / Format

Method	Access Stability	Access Difficulty	Parsing Complexity	Common raw formats	
Manual	М	L	L-H???		
Interactive sessions	L	Н	M-H	HTML elements	
HTTP page requests	М	L	M-H	HTML elements	
API queries	Н	L-M	L	JSON, CSV	

Live Demo

- HTML elements
- API responses

•••	🍅 <mark> y4</mark> Yahoo	Financ	ce - Stock	K Market Li≯	题 Buy Pinot Noir Wine Onli	ne [Unb×	+			~
$\leftarrow \ \rightarrow \ \mathbf{G}$	۵	0	占 ht	tps://www	.danmurphys.com.au/red-v	/ine/pinot-	noir			
Pic	:k up: Bulleen		🗘 Inspe	ector D 🕻	Console 🕞 Debugger 📬 Net	work {}s	tyle Edi	tor 🕜 Per	forma	nance 🕼 Memory 🖻 Storage 뷲 Accessibility ≫ 🛛 🏮 18 🔂 🗍 🚥 🗡
P	Open till 9pm Cart		🗑 Filter L			+	Q 6		ИL С	CSS JS XHR Fonts Images Media WS Other 🗌 Disable Cache No Throttling 🖨 🛱
😤 Filter	1↓ Sort: best match	Sta	Method	Domain	File	Initiator	T 🔻	Transferred	Si	Im Headers Cookies Request Response Timings Stack Trace Security
		200	OPTIO	api.ed	SddPrompt	fetch	plain	739 B	0 B	B
\$19.99 each	\$25.99 each	200	GET	api.da	Keywords?path=/red-wine/pinot-no	i polyfills.67	. json	1.18 kB	8	JSON Raw O
\$18.90 in any six	\$24.70 in any six	200	GET	🔒 api.da	Keywords?path=/red-wine/pinot-no	i polyfills.67	. json	1.18 kB	8	
		200	POST	🔒 api.da	SponsoredAds	polyfills.67.	. json	8.95 kB	3	. SeoMetaTags: Object { Title: "Buy Wine, Beer, Spirits online at Dan Murphy's Alcohol Delivery &
		200	GET	🔒 api.da	Banners?Location=/red-wine/pinot-	polyfills.67.	. json	946 B	3	Bottle Shop", MetaDescription: "Now with contactless delivery, shop online to get drinks delivered to your door or pick up in-store in 30 minutes. Lowest Liguor Price
		200	POST	🔒 api.da	Browse 🔶	polyfills.67	. json	41.21 kB	2	Guarantee. Biggest Range.", NoIndex: false, }
		200	GET	🔒 api.da	Banners?Location=/&Channel=Web	polyfills.67	. json	946 B	3	▼ Bundles: [{}, {}, {}, {}, {}, {}, {}, {}, {}, {}, {}, {}, {}
		200	GET	🔒 aem.d	pinot-noir.model.json	polyfills.67.	. json	4.80 kB		✓ 0: Object { Name: " Little Giant Pinot Noir ", PackDetaultStockCode: "14058/", PackParentStockCode: "140587", }
★★★★★ (48)	***** (0)	200	GET	🔒 api.da	221057,322002	polyfills.67.	. json	4.44 kB		Name: " Little Giant Pinot Noir "
Devil's Corner	Jacob's Creek	200	POST	🔒 api.da	new	polyfills.67.	. json	15.77 kB	8	PackDefaultStockCode: "140587"
Pinot Noir	Classic Pinot Noir	200	POST	🔒 api.da	member offers	polyfills.67	. json	31.64 kB		PackParentStockCode: "140587"
		200	GET	🔒 api.ed	murphy	[:1 (fetch)	json	1.05 kB		Products: [{ }]
		200	GET	🔒 api.ed	SddPrompt	<u>/:1</u> (fetch)	json	946 B	6	·· IsInDefaultList: false
GLUTEN		200	POST	🔒 api.da	SeoMetatags	polyfills.67	. json	1.55 kB	9	IsPersonalised: false
FREE	FREE	200	GET	🔒 aem.d	master.model.json	polyfills.67	. json	3.69 kB		1: Object { Name: "Yarra View Yarra Valley Pinot br>Noir ", PackDefaultStockCode: "79190",
		200	GET	🔒 auth.d	openid-configuration	polyfills.67	. json	2.46 kB		 PackParentStockCode: "9190", } 2: Object { Name: " Devil's Corner Pinot Noir ". PackDefaultStockCode: "902552".
		200	GET	🔒 api.da	Banners?Location=/&Channel=Web	polyfills.67	. json	946 B		PackParentStockCode: "902552", }
Quali Carner	CONSCIENCE CONSCIENCE	200	GET	🔒 api.da	meganav	polyfills.67	. json	52.30 kB		3: Object { Name: " Jacob's Creek Classic Pinot Noir ", PackDefaultStockCode: "328206", DackDecentStockCode: "328206",
		200	GET	🔒 api.da	Categories?Level=0&Location=meg	polyfills.67.	. json	233.86 kB		PackParentSlockCode: 326206 , } 4: Object { Name: " Mud House Claim 431 Vinevard Pinot Noir ". PackDefaultStockCode:
		200	GET	🔒 api.da	Trolley?summary=true&IncludeOrde	polyfills.67	. json	1.42 kB		"917721", PackParentStockCode: "917721", }
\$24.95 each \$24.95 in any six	\$7.95 eac \$7.95 in any six	200	GET	🔒 api.da	Preferences?lsCheckoutV2=true	polyfills.67.	. json	1.94 kB		5: Object { Name: " Cold Snap Cool Climate Pinot >Noir ", PackDefaultStockCode: "171492", DeckDepartStockCode: "171492",
		200	GET	🔒 aem.d	home.model.json	home:184	json	15.86 kB		 PackParentStockCode: "1/1492", } 6: Object { Name: " Elephant In The Room Pinot br>Noir ", PackDefaultStockCode: "608178".
		200	GFT		15 h8c968698c62hb5d4424 is	runtime 00		11 49 kR		PackParentStockCode: "608178", }
		0	317 requ	ests 17.20	6 MB / 4.05 MB transferred Finish:	46.70 s	OMCont	entLoaded: 67	78 ms	IS 7: Object { Name: " Oyster Bay Pinot Noir ", PackDefaultStockCode: "904243",

Final considerations

Ethical & Legal Risk: Retrieval



Ethical & Legal Risk: Data Usage

- It depends, check...
 - Website T&Cs
 - Journal policies
 - robots.txt
 - •



continued acceptance of these terms and conditions.

We may terminate your registered account, at our sole discretion emailing you at the address you have registered stating that the a has terminated.

3. Use of material appearing on the Guardian Site

You acknowledge that, as between Guardian and you, except for content and advertisements (as discussed below), Guardian is th owner or the licensee of all content on the Guardian Site, includi without limitation, all applicable copyrights, patents, trademark secrets, trade names, logos, and other intellectual property right as well as text, images, graphics, logos, audio, video, URLs, meta other material appearing on the Guardian Site ("Guardian Conten Guardian Site and the Guardian Content are protected by the cop laws and other intellectual property laws of the United Kingdom United States and Australia and are protected globally by applica international copyright treaties.

Your use of the Guardian Site and Guardian Content is for your of personal and non-commercial use only. You may download and extracts from the Guardian Content for your own personal and noncommercial use only, provided you maintain and abide by any author attribution, copyright or trademark notice or restriction in any material that you download or print.



Home > Journals > AEA Data and Code Policies and Guidance > AEA Data Legality Policy and Explanations

AEA Data Legality Policy and Explanations

Advice to Authors

A more extensive explanation of the Data Legality Policy's rationale is given below, but a key rationale is that under the Policy the discussion of data legality occurs early in the editorial process so that there are no surprises at the end.

AEA Journal Editors or the AEA Data Editor would be happy to answer questions about data legality before you submit your paper. If you use data that you believe might not be considered legally acquired, please flag this in the online submission portal when submitting your paper.

Starting a Web Scraped Data Project?

Assess Feasibility & Quality:

- Inspect webpages
- Assess consistency
 - How many different webpage architectures?
- Consider data collection vs. analysis unit/resolution
 - Define "research quality" in your context
- Consider website and data owner motives
 - Are there any incentives to obfuscate data?

Resource appropriately:

- For in-house hires, consider experience & knowledge of:
 - Retrieval methods,
 - Scraping etiquette,
 - Parsing methods,
 - Data wrangling
- Consider industry partners if available
- Paid services can be suitable for limited data collection

Appendix

Resources

Web-scraping in R

- Extended tutorial by Hadley Wickham:
 - <u>https://github.com/hadley/web-scraping</u>
- Scraping tools:
 - <u>https://rvest.tidyverse.org</u>
- Polite sessions:
 - <u>https://github.com/dmi3kno/polite</u>

Case Study: Resource Sink

 "Although gathering this massive amount of prices was cheaper online than with traditional methods, it required funding that could not be sustained through grants. Thus, in 2011 we started a company called **PriceStats** that now collects the data and produces high-frequency indexes for central banks and financialsector customers." (Cavallo and Rigobon, 2016, p. 153)

AEA Data Legality Policy

• "A particular concern expressed by many researchers is the treatment of computer-assisted acquisition of data ("scraped data") when such acquisition contravenes terms of use of the data owner. While scraping may contravene terms of use, it may not be illegal. Its legality is not settled under current US law (as of January 2023). Editors will treat papers with scraped data as legally acquired as long as this issue is unsettled when the paper was submitted. Should it in the future become settled law that scraped data is illegal, the AEA will communicate how scraped data will be treated under the Policy."

https://www.aeaweb.org/journals/data/data-legality-policy

References

- Slide 10: Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS), 42–47. <u>https://doi.org/10.1109/CTS.2013.6567202</u>
- Slide 12: Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695–700. <u>https://doi.org/10.1038/s41586-021-04198-4</u>
- Slide 14: Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. Journal of Economic Perspectives, 30(2), 151–178. https://doi.org/10.1257/jep.30.2.151
- Slide 15,26: Foerderer, J. (2023). Should we trust web-scraped data? (arXiv:2308.02231). arXiv. <u>http://arxiv.org/abs/2308.02231</u>
- Slide 24: Nigam, H., & Biswas, P. (2021). Web Scraping: From Tools to Related Legislation and Implementation Using Python. In J. S. Raj, A. M. Iliyasu, R. Bestak, & Z. A. Baig (Eds.), *Innovative Data Communication Technologies and Application* (Vol. 59, pp. 149–164). Springer Singapore. <u>https://doi.org/10.1007/978-981-15-9651-3_13</u>