Learning to Crawl Before You Scrape

Strategies in building a frame through web scraping

Samuel Garber, Mike Gerling & Tyler Wilson

Research and Development

Goal

- Provide an overview of the necessary steps in building a successful survey list frame from open source information (Internet)
  - Today's Example: Industrial Hemp Growers

Why Do This?

- List Building
  - Survey a new commodity
    - Industrial Hemp
  - Enhance a current listing
    - Urban Ag Study (Baltimore, MD)
      - Locate agricultural operations in densely populated areas (cities)
- Adjust for undercoverage
  - https://www.statisticshowto.com/undercoverage-definition/
  - June Area Survey
    - List of all agricultural operations in four states for an undercoverage study
  - Farmers Markets
    - List of all farmers markets across the U.S. for an undercoverage study
- Supplement a Survey's Data or a Survey Replacement

Foundation

- Research and learn the subject matter *(terminology used, regional differences, any special situations)*

- *Industrial Hemp*
  - Uses: Fiber, Grain, Clothing, Paper, Food, Building Materials, etc..
  - Industrial hemp is defined as Cannabis sativa L. and required to be below a THC threshold of 0.3%
    - THC is the main psychoactive compound in cannabis that produces the *high* sensation.

Foundation Cont.

- Define what you want to build a listing of
  - *Industrial Hemp Growers*
    - *Includes/Excludes: Growers, Processors, Laboratories, Seed Growers/Dealers, Transporters, Dispensaries, Medical Marijuana Growers, CBD Growers/Sellers, Sunn Hemp Growers?*
- What information do you want to collect
  - Name of Operation, Operator, Address (street, city, state, zip), Phone(s)
    - Need to link back to your own databases/frame?
      - If yes, what are the key matching variables required?
    - Secondary address? County, Website? Email? Sources? License?
  - *Acres, type of hemp grown, quantity*

Key Words

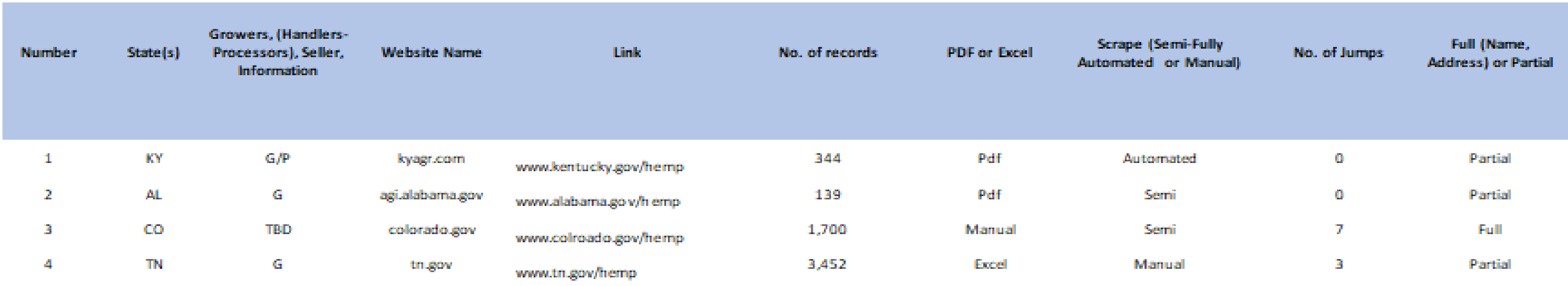
- Develop a list of keywords
  - Industrial Hemp Growers
    - Listing(s), Directory, Associations
      - *State, City*
    - Largest, Biggest
    - *State, City*
    - Correctional Facilities, Churches, Tribal
      - *State, City*
- Too wide of a blanket of keywords will return unwanted entities
  - Example is "Dispensaries" – was also returning eye-glass businesses since they dispense eye-wear

Web Crawling

- Finding the Websites to Scrape the Information from
  - Manual Example: Google – " Industrial hemp growers listing"
    - Website:
      - https://agriculture.ny.gov/system/files/documents/2021/02/authorized_research_partners_0.pdf
- Exploring automated approaches (crawlers)

Website Database

- Create a database of websites to be scraped
  - Number of operations listed, full names and addresses, ease of scraping (pdf, Excel, etc.)
  - Number of Jumps
    - Determine the most cost-effective sites to scrape



Number	State(s)	Growers, (Handlers-Processors), Seller, Information	Website Name	Link	No. of records	PDF or Excel	Scrape (Semi-Fully Automated or Manual)	No. of Jumps	Full (Name, Address) or Partial
1	KY	G/P	kyagr.com	www.kentucky.gov/hemp	344	Pdf	Automated	0	Partial
2	AL	G	agi.alabama.gov	www.alabama.gov/hemp	139	Pdf	Semi	0	Partial
3	CO	TBD	colorado.gov	www.colroado.gov/hemp	1,700	Manual	Semi	7	Full
4	TN	G	tn.gov	www.tn.gov/hemp	3,452	Excel	Manual	3	Partial

Website Database Cont.

- Multiple data formats while scraping
  - Excel
  - Pdf
    - copy and paste into Excel (Adobe Pro option)
  - Website listing with multiple operations per pages and multiple pages
    - Requires substantial manual intervention depending on layout
  - Websites listing less than 5 operations
    - May require substantial manual intervention
      - resources (time vs quantity vs quality)

Web Scraped Database

- Determine your database layout and software employed
  - Initial Industrial Hemp Growers database is developed in Excel

No	State Counter	Duplicate	Match: n (name), o (operation), a (address)	Notes	Name	Name 2	Operation Name	License	Number of Licenses	Address	City	State	Zip	County	Email	Website	Phone1	Phone2	Phone3	Acres	Type	Source 1	Source 2	Source 3	Source 4

Mocked “Real” World Example - Automated or Manual Scrape

Trye Storm
Hemp R Us
32 Wild Air Street
Suite 32
Lake, NY 12019
518-777-0202

Kathy Wind
We R Hemp - Be Sweet Inc
PO Box 40
Albany, NY 12022
Lake, NY 12019
518-777-0202 ext 113
emai: wrhemp@google.com
Charles Drough
Watertown, NY 12334

Lisa Swift
14 High Street
Cascades, NY www.swift.com

* For additional Information please see our website at www.hempusa.com

Geroge Mika
777-243-7827
www.accesshemp.com

Hillside Farms LLC
Troy Awash
PO Box 34, 577 Water Drive
Watertown, NY 14457
518-234-8977
518-344-255 Ext 101
troyawah@yahoo.com
www.amazing hemp.com

Database record’s information needs to be horizontal

Web Scraping Techniques

- Does the source copy/import into Excel well? (or) places everything into one cell (double ugh!)

- Excel Techniques Employed
  - Transpose (vertical to horizontal)
  - Formulas – copy and repeat
  - Remove all blank rows
  - Remove images
  - Conditional formatting (address, city, state, zip)

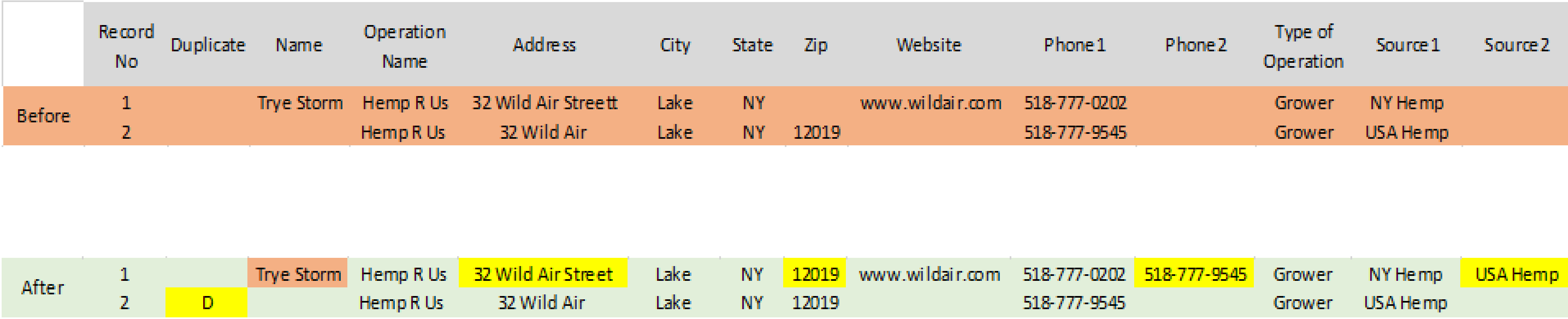
Automated Web Scraping

- Websites are blocking these efforts
- Requires multiple virtual machines with multiple IP addresses
  - Waiting period to repeat next scrape attempt

After Scraped Database is Created Next Step is Data Enhancing

- Duplication Marking/Removal (First Round)
  - Excel's Sorting and Filter functions can zero in on the duplicate records and collapse information between records
    - Lessens the number to enhance and clean later on in the process

Example of Duplication Marking and Combining Data into One Record



	Record No	Duplicate	Name	Operation Name	Address	City	State	Zip	Website	Phone 1	Phone 2	Type of Operation	Source 1	Source 2
Before	1		Trye Storm	Hemp R Us	32 Wild Air Streett	Lake	NY		www.wildair.com	518-777-0202		Grower	NY Hemp	
	2			Hemp R Us	32 Wild Air	Lake	NY	12019		518-777-9545		Grower	USA Hemp	
After	1		Trye Storm	Hemp R Us	32 Wild Air Street	Lake	NY	12019	www.wildair.com	518-777-0202	518-777-9545	Grower	NY Hemp	USA Hemp
	2	D		Hemp R Us	32 Wild Air	Lake	NY	12019		518-777-9545		Grower	USA Hemp	

Data Enhancing Cont.

- Address and phone lookups (email, websites, etc.)
  - Thomson Reuters CLEAR investigation software
  - SAS JMP Geocoder via Google Maps
  - Google
  - White Pages, Yellow Pages, Manta
  - SAS programs that check zip codes against US Postal Service’s for the particular city and state

  - Balance between Quality and Quantity
    - What is the volume of the data?
      - 1,000 records vs 10,000 vs 100,000, 1 million records

Data Cleaning

- Duplication Removal
  - Use Excel (again) to zero in on any additional duplicate records
  - We find this partial automated approach is faster if we need to look up additional information to decide if truly a duplicate record or not
    - Example: Two operations can be different even though their address is the same
      - One is an industrial hemp **grower**
      - The other is an industrial hemp **processor**

Data Cleaning Cont.

- Import into our Legacy List Frame Systems and SAS Programs
  - Developed by our Frames Maintenance Group
    - St Louis, MO
      - SAS (Functions: Compress, Tranwrd, Compbl, Scan, various Macros)
  - Removal of weird characters, symbols, etc.
  - Final Duplication removal (looking into fuzzy matching)
  - Cleans up address fields
  - Checks city with state and zip
  - Email formatting
  - Phone formatting

Results: Seven hemp-producing States, 101 websites found, total of 9,981 hemp producers web scraped after data cleaning

Below is a breakdown of these numbers

State	Operations Scraped	Operations Remaining After Removing Duplicates	Percentage Remaining
Colorado	5,212	2,684	51%
Illinois	820	702	86%
Missouri	865	527	61%
Montana	722	659	91%
Nevada	540	375	69%
New York	1,376	746	54%
Tennessee	5,740	4,288	75%
Total	15,275	9,981	65%

Articles & Publications

- Blazquez, D., Domenech, J., Gil, J.A., and Pont, Ana. (2019). Monitoring e-commerce adoption from online data. Knowledge Information Systems, 60, 227–245. DOI: https://doi.org/10.1007/s10115-018-1233-7.

- de Pedraza, P., Visintin, S., Tijdens, K., Kismihók, G. "Survey vs Scraped Data: Comparing Time Series Properties of Web and Survey Vacancy Data", IZA Journal of Labor Economics , Volume 8: Issue 1,
DOI: https://doi.org/10.2478/izajole-2019-0004.

- Cavallo, A., and R. Rigobon, (2016), "The Billion Prices Project: Using Online Research for Measurement or Research," Journal of Economic Perspectives, 31(2), 151-178.

- Chow, T. E., Y. Lin, and W. D. Chan, (2011), "The Development of a Web-based Demographic Data Extraction Tool for Population Monitoring," Transactions in GIS, 15(4), 479-494. DOI: https://doi.org/10.1111/j.1467-9671.2011.01274.x.

Articles & Publications

- Rhodes, B. B., A. F. Kim, and B. R. Loomis, (2015), “Vaping the Web: Crowdsourcing and Web Scraping for Establishment Survey Frame Generation,” In Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference, available at https://fcsm.sites.usa.gov/files/2016/03/H3_Rhodes_2015FCSM.pdf.

- Vargiu, E. and Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artificial Intelligence Research, 2(1), 44-54, DOI: https://doi.org/10.5430/air.v2n1p44.

- Webb, L. M., D. M. Gibson, Y. Wang, H. C. Chang, and M. Thompson-Hayes, (2015), “Selecting, Scraping, and Sampling Big Data Sets from the Internet: Fan Blogs as Exemplar,” SAGE Research Methods Case, London: SAGE Publications, Ltd., available at https://pdfs.semanticscholar.org/48f8/4535a0607c7e64f87c61faf176ef3c5161fc.pdf?_ga=2.244060101.1587555821.1595364779-1428202309.1595364779.

- Young, L., Hyman, M., Rater B.. (2018), “Exploring a Big Data Approach, to Building a List Frame for Urban Agriculture: A Pilot Study in the City of Baltimore”,. Journal of Official Statistics, Vol. 34, No. 2, 2018, pp. 323–340, http://dx.doi.org/10.2478/JOS-2018-0015.

Additional Contributors and Thank You’s

- Linda J. Young, Ph.D.
- Denise Abreu
- Kara Daniel
- Jessica Winterowd
- Kay Turner
- Mike Bellow, Ph.D.

Samuel Garber	samuel.garber@usda.gov	202-692-0284
Michael Gerling	michael.gerling@usda.gov	202-692-0277
Tyler Wilson	tyler.wilson@usda.gov	202-692-0290