Advancing Privacy Research: A Novel Realistic Persona-Based Dataset

Carson Godwin[1,], Wen-Chen Hu[2,†], Hosam Alamleh[1,†], Ali Abdullah S. AlQahtani[3,†], and AbdElRahman ElSaid[1,*,**,†]**

[1]Department of Computer Science, University of North Carolina Wilmington, Wilmington 28403
[2]Department of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks 58203
[3]Department of Computer Systems Technology, North Carolina Agricultural and Technical State University, Greensboro, NC 27411
[*]Godwin's advisor on the project
[**]Corresponding authors: Carson Godwin (crg9100@uncw.edu), AbdElRahman ElSaid (elsaida@uncw.edu)
[†]these authors contributed equally to this work

ABSTRACT

We introduce a unique approach to privacy research by creating a virtual persona that mimics human web-searching behaviors. The persona's activities, categorized into 'morning', 'afternoon', and 'evening', were automated using the Selenium WebDriver, enabling the persona to conduct searches as a real user would. The resulting dataset comprises 1,537 records, each representing a unique search query. Each record contains the first two pages of a query result, including the query keywords and a list of the first 2 pages of the query result. The study offers a fresh perspective on the study of privacy and personalization in online environments. The potential for reusing this dataset is significant, as it can be applied to studies on privacy, data collection, and search engine personalization, and it can be used to develop and test algorithms and models that aim to protect user privacy. Furthermore, this work not only produces a dataset but also establishes the framework for generating additional similar datasets that can serve the purpose of conducting more comprehensive research on search engines.

Background & Summary

In the evolving field of privacy research, unique methodologies has been employed to protect users' privacy[1–3]. A persona, a virtual identity, was created to mimic human web-searching behaviors. This persona's activities were categorized into 'morning', 'afternoon', and 'evening', reflecting the different times of the day. The Selenium WebDriver, a tool for automating web tasks[4], was used to enable the persona to conduct searches as a real user would. This approach allowed the search engine to learn about the persona and customize the search results accordingly.

The data[*] generated is a collection of 1,537 records, each representing a unique search query. Each record contains the first two pages of a query result, including the query keywords and a list of the first 2 pages of the query result. Each list includes the title of the result and the link to the page of the result and its sub-page.

This study is rooted in the broader discourse on privacy and data collection. It draws on various studies on privacy concerns[5], fair information practices[6], and the balance between privacy protection and data collection[7,8]. The study contributes to the ongoing discussion on privacy in the digital age and the impact of big data and analytics on privacy and user control.

The creation of this dataset was motivated by the desire to advance privacy research. The dataset offers a novel approach to data collection that mimics human behavior and allows for the study of personalized search engine results. The potential for reusing this dataset is significant. It can be applied to studies on privacy, data collection, and search engine personalization, and it can be used to develop and test algorithms and models that aim to protect user privacy.

A schematic overview depicts the study design, providing a visual representation of the persona profile and the data collection process. A sample of the results demonstrates the change in results for the same query keywords when queried over different periods of time.

This work introduces a unique, personalized web-searching dataset, a significant contribution to various scientific disciplines. To the best of our knowledge, this dataset is the first of its kind to be openly published, marking a significant milestone in the field. The novelty of this contribution is rooted in two key factors. Firstly, the sensitivity associated with acquiring such a dataset is considerable. The collection of this type of information often severely infringes upon the privacy of individuals, exposing them to potential malicious attacks and threats. However, this dataset has been carefully curated to ensure that it

[*]Harvard Dataverse: https://doi.org/10.7910/DVN/GOHBTR

Figure 1. ***Persona Profile***[†]

does not violate or risk any personal privacy. This careful approach to data collection sets a new standard for research in this area. Secondly, the dataset provides a detailed, realistic interaction between the persona and the search engine. This offers invaluable insights for researchers interested in studying and investigating how the search engine responds to user queries to offer a personalized experience. The depth of interaction captured in this dataset provides a rich resource for further study. In addition, the persona and the program developed to collect the data are made available as open-source. This will allow researchers to expand and enrich the database with more queries and other personas that reflect different personalities and characteristics. Looking ahead, our team plans to continue expanding the dataset by adding more queries and incorporating other personas. Our research plans include investigating web users' privacy vulnerabilities and exploring ways to preserve them. We will continue to publish the tool's code on our code-repository, and the data updates and expansion on our data-repository, ensuring that our work remains accessible to the wider research community.

Methods

On the one hand, collecting real person data violates the privacy of the person and expose them to public, thus collecting the search history of real search engine user was off limits. On the other hand, synthetic data is susceptible to separation from reality and the inconsistency of the data elements and features. The solution was to carefully establish a persona that represent the interests and behavior of a typical member of our contemporary world, within given specific socioeconomic boundaries. Figure 1 illustrates traits of the established persona. The data was collected by building a search record over time for the created persona.

Creating the cyber history of the a persona through web scraping involved a number of steps and considerations. First, we implemented the web scraper with a personalized approach in mind; we wanted it to exhibit behaviors like a human would in different times of the day, avoiding being flagged by the search engine as non-real person. Avoiding being detected as non-human was done setting time gaps between each query submitted to the search engine. Also, to maintain the affinity to human behavior, we organized our search queries into different categories according to *'morning'*, *'afternoon'*, and *'evening'*.

[†]Picture generated from unrealperson.com

Algorithm 1 Data Scrapper

```
keywords_by_time ← {
morning_lists : [morning_food, morning_news, religion_paractice, random, sports]
afternoon_lists : [afternoon_interests, sports, afternoon_food, religion_paractice]
evening_lists : [evening_interests, sports, evening_food, religion_paractice]
}
procedure GetTrait
    ▷ Get Personality trait list
    if now.hour ≥ 5 and now.hour < 12 then
        time_of_day =' morning
    else if now.hour ≥ 12 and now.hour < 18 then
        time_of_day = afternoon
    else
        time_of_day = evening
    trait ← keywords_by_time[time_of_day]
    return trait
procedure GetSearchResults
    results ←
    for query ← 1 ... query_num do
        trait ← GetTrait()
        for page ← 1 ... page_num do
            search ← page_links_headers
        results[query] ← search
    return results
procedure Main
    Login to User Account
    results = GetSearchResults()
    StoreResults(results)
```

These categories were saved in respective text files, each containing numerous queries' keywords that were deemed relevant for the particular time of day.

In the Python script [‡], we used the Selenium WebDriver [§] for automating the tasks. The first step was to log into Google, allowing our script to conduct searches just as a real user would through their personal account. Creating a personal account for the persona and performing the queries through it allows the search engine to learn about the persona and profile it, personalizing the results to fit the interests of cyber-behavior of the persona.

Algorithm 1 illustrates the pseudo-code of the tool designed for collecting the data. The actual search process was organized in a function called *GetSearchResults*. In this function, the script selected a category based on the current time of the day, selected a random query from the chosen category, and then carried out the Google search. To ensure a good mix of results, the script performed five searches, each time with a random query. The search results from each query were then stored in a dictionary (hashmap), with the key being the search query and the value being a list of the text from the search results. The script was designed to scrape only the first two pages of each Google search, giving a total of 20 results per search query. We chose this limit to balance between data quantity and the risk of overloading the server with requests. After all the searches were conducted, the script saved the results dictionary into a text file using the *StoreResults* function. The structure of this text file was straightforward; each line contained a key-value pair from the dictionary, with the key and value separated by a colon.

Through this process, we successfully created a persona that exhibited unique web-searching behaviors based on time of the day, and collected a sizable amount of data reflecting these behaviors. The data is published on our data-repository[9].

Figure 2. ***A Record From the Database***

[‡]Code GitHub repository: https://github.com/CarsonGodwin/ML-Webscraper.git
[§]https://github.com/SeleniumHQ/selenium/

Data Records

The dataset comprises 1,537 records, each record represents a search query. As described in the Methods Section, each data record contains the first two pages of a query result. The query keywords are at the beginning of the record followed by a colon ':', followed by a list (bounded by square parenthesis: [and]) of the first 2 pages of the query result. The lists of the results in each record contain the title of the result and the link to the page of the result and its sub-page (Figure 2). Table 1 depicts a sample of the results with queries keywords in the first column, the title of the results in the second column, and the URL of the results in the third column. Table 2 shows that the data exhibits change in the results for the same query keywords when queried over different periods of time, which will be discussed in the next section.

Query	**Result**	
	TITLES	URLs
How to start a garden	• How to make a garden: a beginner's guide : The Salt : Life KitNPR	www.npr.org
	• How to start your own vegetable garden for the New Yearhappysprout.com	www.happysprout.com
	• How To Start A Garden from Scratch Without Breaking the Bankgardentherapy.ca	gardentherapy.ca
	• Beginning Vegetable Garden Basics: Site Selection and Soil Preparationwisc.edu	hort.extension.wisc.edu
	• 10 Easy Steps to Create Gardens in Your Yard for the First ...Better Homes and Gardens	www.bhg.com
	• How to Start a Garden – 10 Steps to Gardening for BeginnersCommon Sense Home	commonsensehome.com
	• How to Start and Plan a Garden in 14 StepsThe Spruce	www.thespruce.com
	• How to Start a Backyard Garden: 11 Steps for New ...MasterClass	www.masterclass.com
	• Gardening for Beginners How to Start a Garden in 8 Simple ...Growing In The Garden	growinginthegarden.com

Table 1. ***Results Sample***

Query	**Tiltes**
1	• Gen Z is the most pro union generation alive. But will they ...NPR • Prounion Definition & Meaning - Dictionary.comdictionary.com • Gen Z is the most pro union generation alive. Will they ... - NPRNPR • The Union Advantage | U.S. Department of LaborU.S. Department of Labor (.gov) • "Tesla Factory Announces Union Bid, Testing Elon Musk's Very Public ...vanityfair.com • What Is the PRO Act?AFL–CIO • Why Gen Z is the most pro-union generation - MarketplaceMarketplace • PRO Union ConsultingPRO Union Consultinghttp • The Union AdvantageU.S. Department of Labor (.gov) • Senators Debate Pro-Union LegislationSociety for Human Resource Management • Labor Unions During the Great Depression and New DealLibrary of Congress (.gov) • "President Biden: "The Most Pro-Union President You've ...Miller Johnson
2	• As Pro-Union Sentiment Reaches a Fifty-Year High, U.S. ...The New Yorker • Prounion Definition & Meaning - Dictionary.comdictionary.com • News & Commentary: April 13, 2023 OnLaboronlabor.org • The Union Advantage | U.S. Department of LaborU.S. Department of Labor (.gov) • "Tesla Factory Announces Union Bid, Testing Elon Musk's Very Public ...vanityfair.com • What Is the PRO Act?AFL–CIO • Why Gen Z is the most pro-union generation - MarketplaceMarketplace • PRO Union ConsultingPRO Union Consultinghttp • Senators Debate Pro-Union LegislationSociety for Human Resource Management • The Union AdvantageU.S. Department of Labor (.gov) • Labor Unions During the Great Depression and New DealLibrary of Congress (.gov) • RELEASE: Gen Z Is the Most Pro-Union Generation, New ...Center for American Progress
3	• As Pro-Union Sentiment Reaches a Fifty-Year High, U.S. ...The New Yorker • Prounion Definition & Meaning - Dictionary.comdictionary.com • The Union Advantage | U.S. Department of LaborU.S. Department of Labor (.gov) • "Tesla Factory Announces Union Bid, Testing Elon Musk's Very Public ...vanityfair.com • What Is the PRO Act?AFL–CIO • Why Gen Z is the most pro-union generation - MarketplaceMarketplace • "You may have heard of the 'union boom.' The numbers tell ...NPR • The Union AdvantageU.S. Department of Labor (.gov) • PRO Union ConsultingPRO Union Consultinghttp

	• Labor Unions During the Great Depression and New DealLibrary of Congress (.gov) • Pro-Unions WebPro-Unions Web
4	• As Pro-Union Sentiment Reaches a Fifty-Year High, U.S. ...The New Yorker • Prounion Definition & Meaning - Dictionary.comdictionary.com • The Union Advantage	U.S. Department of Labordol.gov • Labor Union - Definition, Explained, History, Types, Exampleswallstreetmojo.com • What are the pros and cons of being in a union? - Zippiazippia.com • What Is the PRO Act?AFL–CIO • "Saunders: President Biden is 'most pro-union, pro-worker ...American Federation of State, County and Municipal Employees • Pro-Unions WebPro-Unions Web • PRO Union ConsultingPRO Union Consultinghttp • "How the PRO Act restores workers' right to unionizeEconomic Policy Institute • Pro-labor? Biden aims to prove it with unionized 2024 staffAP News

Table 2. ***Search Engine Responses Sentiments Change*** **– keyword: `pro union`**

Technical Validation

Comparing the data to real people profiles on the search engine would compromise the users' privacy, which defies to purpose of this work. In addition, no database about real people web-browsing is publicly available for ethical, commercial, and legal reasons. This dataset is proposed as a medium to analyse the search engines attitude against users' queries, rather than a medium to analyse human behavior when interacting with a 'smart' search engine. Therefore, the way to technically validate this data is by ensuring that it achieve it's intended goal: to study search engines behavior. This can be tested by ensuring that this data will cause the search engine to respond differently to queries as user profile is being built based on the search history. Analysing the data from this prospective shows the search engine changing behavior as the submission of queries progresses through time.

The data in Table 2 in previous Section shows that the results gradually shift from revolving around *'Gen Z'* involvement with *'trade unions'*(since the persona is in their early 20's) to results about unions definition, unions politics, unions as a social labor movement, and unions organization. This manifests that the search engine model started to catch up with the interests of the persona based on their search history, which was built up through queries submission over time.

Examining the outcomes of query responses reveals the adaptive nature of the search engine's behavior. Among the top 12 search results, comparing the first and the second queries indicates that 8 search results remained consistent but in altered sequences, while 4 search results changed. Notably, when contrasting the first and fourth queries, only 4 search results remained unchanged, while 8 were changed, each time with distinct arrangements. In addition, the words *'Generation'* and *'Gen Z'* appeared 3 times in the first result, twice in the top 3 links, appeared 2 times in the second result, none in the top 6 links, one time in the third result, and it disappeared from the fourth result. This underscores how the constructed persona elicited varied responses from the search engine.

Code Availability

This work not only produces a dataset but also establishes the framework for generating additional similar datasets that can used to conduct extensive research on search engines. The code used to generate the dataset is published as open-source on the Code GitHub repository: https://github.com/CarsonGodwin/ML-Webscraper.git. The code is written in Python and its backbone module is the "Selenium WebDriver" module.

Author Contributions

C.G. developed the code, conceived the program runs, and wrote the Methods section. W.H. reviewed the novelty of the collected data and the authenticity of the idea. H.A., A.A., and A.E. wrote the other paper sections, and reviewed the work and the data. All the authors reviewed the manuscript. The team is working on internet user-privacy research problems.

Competing Interests

The authors declare **no** competing financial and/or non-financial interests in relation to the work described.

References

1. Xiong, J. *et al.* A personalized privacy protection framework for mobile crowdsensing in iiot. *IEEE Transactions on Ind. Informatics* **16**, 4231–4241 (2019).

2. Qi, L. *et al.* Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment. *IEEE Transactions on Ind. Informatics* **17**, 4159–4167 (2020).

3. Gai, K., Wu, Y., Zhu, L., Qiu, M. & Shen, M. Privacy-preserving energy trading using consortium blockchain in smart grid. *IEEE Transactions on Ind. Informatics* **15**, 3548–3558 (2019).

4. Gundecha, U. & Avasarala, S. *Selenium webdriver 3 practical guide: End-to-end automation testing for web and mobile browsers with selenium webdriver* (Packt Publishing Ltd, 2018).

5. Cheah, J.-H., Lim, X.-J., Ting, H., Liu, Y. & Quach, S. Are privacy concerns still relevant? revisiting consumer behaviour in omnichannel retailing. *J. Retail. Consumer Serv.* **65**, 102242 (2022).

6. Gellman, R. Fair information practices: A basic history-version 2.22. *Available at SSRN* (2022).

7. Saura, J. R., Ribeiro-Soriano, D. & Palacios-Marqués, D. Assessing behavioral data science privacy issues in government artificial intelligence deployment. *Gov. Inf. Q.* **39**, 101679 (2022).

8. Rustambekov, I., Safoeva, S., Rodionov, A. & Uktam, R. Balance between data collection and privacy in the context of smart cities. *Int. J. Cyber Law* **1** (2023).

9. ElSaid, A. Replication Data for: Advancing Privacy Research: A Novel Realistic1 Persona-Based Datase, 10.7910/DVN/GOHBTR (2023).