Tutorial

Generating Contextual Variables From Web-Based Data for Health Research: Tutorial on Web Scraping, Text Mining, and Spatial Overlay Analysis

Pablo Galvez-Hernandez[1,2], PhD; Angelina Gonzalez-Viana[3], PhD; Luis Gonzalez-de Paz[4,5], PhD; Ketan Shankardass[6,7], PhD; Carles Muntaner[1,8], PhD

[1]Lawrence S Bloomberg Faculty of Nursing, University of Toronto, Toronto, ON, Canada

[2]Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[3]Public Health Agency of Catalonia, Health Department, Barcelona, Spain

[4]Primary Healthcare Transversal Research Group, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain

[5]Consorci d'Atenció Primària de Salut Barcelona Esquerra, Barcelona, Spain

[6]Department of Heath Sciences, Wilfrid Laurier University, Waterloo, ON, Canada

[7]MAP Centre for Urban Health Solutions, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, ON, Canada

[8]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Pablo Galvez-Hernandez, PhD
Institute of Health Policy, Management and Evaluation
Dalla Lana School of Public Health
University of Toronto
Health Sciences Building, 4th Fl.
155 College St
Toronto, ON, M5T 3M6
Canada
Phone: 1 6475752195
Email: pau.galvez@utoronto.ca

Abstract

Background: Contextual variables that capture the characteristics of delimited geographic or jurisdictional areas are vital for health and social research. However, obtaining data sets with contextual-level data can be challenging in the absence of monitoring systems or public census data.

Objective: We describe and implement an 8-step method that combines web scraping, text mining, and spatial overlay analysis (WeTMS) to transform extensive text data from government websites into analyzable data sets containing contextual data for jurisdictional areas.

Methods: This tutorial describes the method and provides resources for its application by health and social researchers. We used this method to create data sets of health assets aimed at enhancing older adults' social connections (eg, activities and resources such as walking groups and senior clubs) across the 374 health jurisdictions in Catalonia from 2015 to 2022. These assets are registered on a web-based government platform by local stakeholders from various health and nonhealth organizations as part of a national public health program. Steps 1 to 3 involved defining the variables of interest, identifying data sources, and using Python to extract information from 50,000 websites linked to the platform. Steps 4 to 6 comprised preprocessing the scraped text, defining new variables to classify health assets based on social connection constructs, analyzing word frequencies in titles and descriptions of the assets, creating topic-specific dictionaries, implementing a rule-based classifier in R, and verifying the results. Steps 7 and 8 integrate the spatial overlay analysis to determine the geographic location of each asset. We conducted a descriptive analysis of the data sets to report the characteristics of the assets identified and the patterns of asset registrations across areas.

Results: We identified and extracted data from 17,305 websites describing health assets. The titles and descriptions of the activities and resources contained 12,560 and 7301 unique words, respectively. After applying our classifier and spatial analysis algorithm, we generated 2 data sets containing 9546 health assets (5022 activities and 4524 resources) with the potential to enhance social connections among older adults. Stakeholders from 318 health jurisdictions registered identified assets on the platform between July 2015 and December 2022. The agreement rate between the classification algorithm and verified data sets ranged from 62.02% to 99.47% across variables. Leisure and skill development activities were the most prevalent (1844/5022, 36.72%). Leisure and cultural associations, such as social clubs for older adults, were the most common resources (878/4524, 19.41%). Health asset registration varied across areas, ranging between 0 and 263 activities and 0 and 265 resources.

Conclusions: The sequential use of WeTMS offers a robust method for generating data sets containing contextual-level variables from internet text data. This study can guide health and social researchers in efficiently generating ready-to-analyze data sets containing contextual variables.

(JMIR Public Health Surveill 2024;10:e50379) doi: 10.2196/50379

KEYWORDS

web scraping; text mining; spatial overlay analysis; program evaluation; social environment; contextual variables; health assets; social connection; multilevel analysis; health services research

Introduction

Background

Contextual variables refer to the social or physical attributes of geographic or jurisdictional areas (eg, country, city, neighborhood, and administrative health area) that are not derived from the characteristics of their members [1]. Common examples include social cohesion [2], social capital [3], and presence of green spaces [4]. Contextual variables have multiple applications in health and social research. As people living in the same community or context are likely to be exposed to a similar environment, contextual variables can be used in multilevel models to explain variability in health outcomes [5].

Although information on some contextual variables, such as census data, is widely available, accessing context-level data in emerging research fields can pose significant challenges. For example, monitoring systems may not exist yet to fully capture the social determinants of health (SDOH) across delimited areas. In addition, there may not be data available on the exposure and implementation of large-scale interventions targeting SDOH, making program and implementation evaluation studies challenging or impossible [6]. This could be the case for regional or state public policies and public health programs, such as provincial public health programs to promote local intersectoral collaborations to tackle SDOH [7] or national legislation to promote healthy nutrition to prevent obesity [8]. As these policies and programs can be implemented without an evaluation plan, and data might be complex or unavailable, they often remain unevaluated [9].

When structured databases or primary data gathering are not feasible, the internet can be a valuable resource for compiling information to define contextual variables. However, this presents several challenges: internet data are often cluttered, fragmented, and spread over multiple websites [10]. Moreover, the content of most websites is not designed for use by health researchers nor is it grouped by relevant contextual areas. To overcome these challenges, we developed a novel 8-step method, which we have termed web scraping, text mining, and spatial overlay analysis (WeTMS) to collect large amounts of internet data from websites, transforming it into meaningful data sets containing research-relevant variables, and classifying them based on delimited geographical or jurisdictional areas.

This method combines the techniques used in web scraping, text processing and mining, and spatial analysis. Web scraping, also known as web data mining, involves the creation of programs that can automatically download, parse, organize, and store information collected from the web in structured data sets [11]. This process is more efficient and less prone to errors compared with the traditional and laborious process of manually copying and pasting internet information into a spreadsheet [11]. Web scraping has been gaining traction in health research, fueling the rise of *infodemiology*, which analyzes the spread and impact of web-based information to inform public health and policy [12]. As of January 2023, a search of the keyword "web scraping" in Medline yielded 105 records, 95 of which were published starting from 2019. Articles using web scraping in health and social research mostly used information from social media [13,14] (eg, Twitter, Instagram, and TikTok), forums [15,16], business and review websites [17], and news web pages [18].

Similarly, text mining has been increasingly applied in health and social research [19]. Text mining is the process of extracting meaningful information from large volumes of unstructured text data using techniques such as text classification, sentiment analysis, and pattern recognition [19]. Examples include using sentiment analysis on social media posts to identify health and mental well-being issues [20] and characterizing mental health problems [21]. In addition, topic modeling has been used to understand public perceptions of the COVID-19 pandemic on Twitter [22] and to uncover health-related topics on social media [23].

Spatial overlay analysis is a group of methodologies used in geographic information systems to simultaneously display multiple layers of spatial information and assess the relationships between different geographic features and attributes [24]. Spatial overlay analysis can be used to examine the relationships between multiple layers of geospatial data to locate spatial points (eg, coordinates) in delimited geographic or jurisdictional areas. Geographic information system methods have been used in health geography and environmental epidemiology to study the geographic incidence or distribution of diseases [25].

When census data or data sets containing contextual variables are unavailable, researchers may have to engage in laborious manual extraction of web-based data, which can be time-consuming and susceptible to inaccuracies [11]. Currently, there is a gap in the literature regarding methods that enable researchers to automatically convert large volumes of internet text information into meaningful, ready-to-analyze data sets containing contextual data. We propose that by combining techniques used in WeTMS, researchers can efficiently extract, process, and geolocate vast amounts of internet text data to produce structured data sets that encompass variables reflecting the contextual characteristics of specific geographic or jurisdictional areas.

Objectives

The aims of this study are 2-fold. First, we outline the implementation of the WeTMS method through a research case, creating data sets with contextual variables on health assets that could improve social connections for older adults across various health jurisdictions in Catalonia, Spain. Second, we analyze these data sets to describe the characteristics and registration trends of these health assets by local stakeholders.

In this tutorial, we first introduce the WeTMS method and describe its application to a research case for compiling data sets of health assets that could enhance social connections among older adults in the health jurisdictions of Catalonia. These assets include activities and resources in the community that can facilitate social interaction, such as social activities, walking groups for retirees, libraries, and senior community centers [26].

Next, we use these new data sets to extract assets with the potential to foster older adults' social connections and conduct a descriptive analysis to explore their characteristics and asset registration trends across jurisdictions. This analysis demonstrates the potential application of this method in program evaluation. In addition, we discuss the challenges that health and social researchers may face during the WeTMS process and provide resources and programming codes to facilitate its application in other areas of research.

Methods

Context and Data Sources

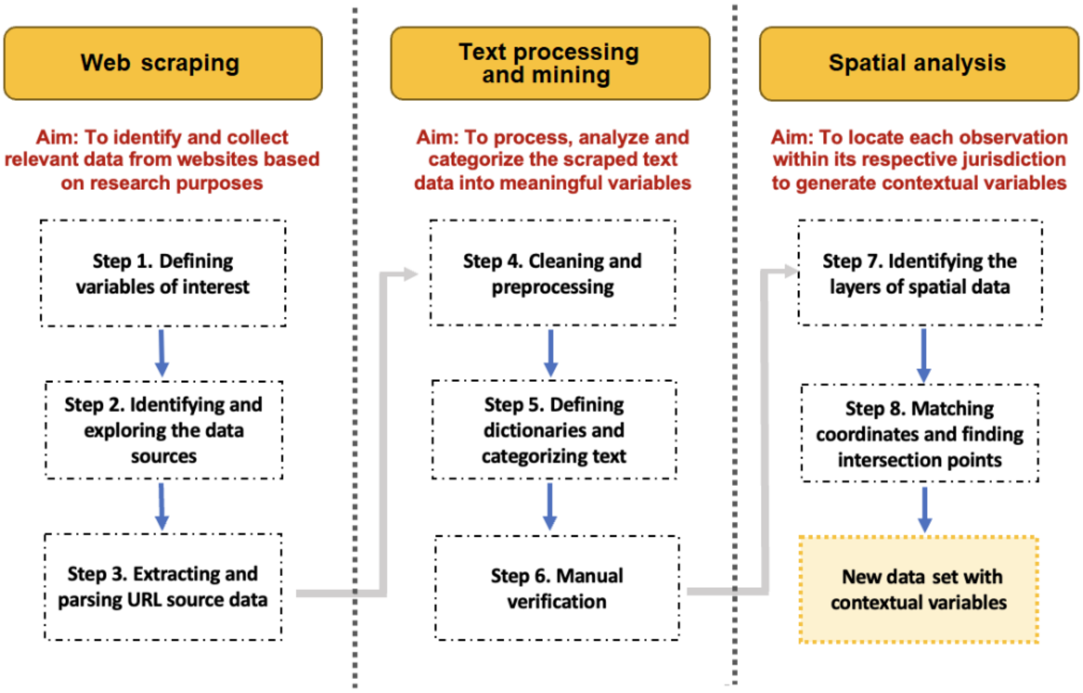
In 2015, the government of Catalonia launched the Assets and Health platform as a component of 2 provincial public health programs that aimed to promote intersectoral collaborations among health and nonhealth organizations to tackle complex public health issues, including older adults' lack of social connections [7]. The Assets and Health platform (created by the Asturias Health Observatory and shared through the Spanish Community Health Alliance) is a search engine and repository where stakeholders from multiple local organizations can register community health assets [27]. Health assets are activities and resources within the community that contribute to maintaining the health and well-being of individuals and groups [28].

Health assets were registered as "activities" (time-bound initiatives and structured interventions, like arts and crafts or supervised walking outings) and "resources" (permanent community fixtures such as associations, parks, and civic centers). Once registered, each health asset is stored on an individual website detailing characteristics, such as its title, description, location, and target population. These individual websites are linked to a search engine, enabling stakeholders to locate assets available in their basic health areas (BHAs), which can be used in collaborative interventions to address public health problems. Each BHA in Catalonia is a local health jurisdiction that functions as an administrative unit within the Catalan healthcare system [29]. In urban settings, BHAs typically cover specific neighborhoods or districts, whereas in rural areas, they may span one or more municipalities, as determined by demographic, epidemiological, and accessibility considerations.

Overview of the WeTMS Method

The 8 steps of the proposed method are summarized in Figure 1. The first 3 steps involve identifying and extracting website data through web scraping, and then storing the information in structured data sets to facilitate their analysis. Steps 4-6 describe the application of text processing and mining techniques to analyze the scraped data, identify patterns in the text content, and classify the data into new variables and categories. Steps 7 and 8 elaborate on the use of spatial overlay analysis to locate data within delimited geographic or jurisdictional areas.

Figure 1. Overview of the web scraping, text mining, and spatial overlay (WeTMS) steps for generating contextual variables from unstructured web-based data.


Web scraping
Text processing and mining
Spatial analysis
Aim: To identify and collect relevant data from websites based on research purposes
Aim: To process, analyze and categorize the scraped text data into meaningful variables
Aim: To locate each observation within its respective jurisdiction to generate contextual variables
Step 1. Defining variables of interest
Step 2. Identifying and exploring the data sources
Step 3. Extracting and parsing URL source data
Step 4. Cleaning and preprocessing
Step 5. Defining dictionaries and categorizing text
Step 6. Manual verification
Step 7. Identifying the layers of spatial data
Step 8. Matching coordinates and finding intersection points
New data set with contextual variables

Steps 1 to 3: Using Web Scraping for Data Extraction

Step 1: Defining Variables of Interest

The target information, including the type of data and the desired outputs for web scraping, was first outlined to avoid extracting irrelevant information. We aimed to generate context-level variables capturing the attributes and registration dates of community health assets to enhance social connections among older adults in Catalan BHAs. The data to be extracted from each website detailing a health asset included text elements such as title, description, target population, location, asset registration date, cost, duration, and activity topics. Other data types that could be targeted for extraction include images, links, and metadata, while outputs might be structured data sets such as CSV files, which contain prespecified variables.

Step 2: Identifying and Exploring the Data Sources

The second step involves identifying the URLs or web addresses containing the target information and understanding how their content is structured. Identifying URLs can present challenges such as information being dispersed across multiple websites or URLs being hidden or changing [30]. Consequently, sites may be missed because the web scraping "crawler" (the portion of code responsible for finding each URL) requires exact web addresses [30].

In our initial exploration of the Assets and Health search engine, only the last 100 registered health assets were displayed, and those URLs were hidden. A pattern in the URLs for websites describing each health asset, comprising a fixed segment and variable reference number, was identified using Chrome DevTools for network inspection.

The source code of the target websites, usually HTML, was examined to discern their organization and structure. The attributes of HTML elements containing target data (eg, asset titles and descriptions) were identified and used to program the web scraper. Additional information and resources required to implement step 2 are available in Multimedia Appendix 1 [11,31-36].

Step 3: Extracting and Parsing URL Source Data

After identifying the relevant URLs and HTML elements, a web scraper comprising a crawler, parser, and data handler was developed using Python 3.10 [37] on the PyCharm 2022.2.2 environment, and the libraries "requests" [38], "beautifulsoup4" [39], and "pandas" [40]. The web scraper also incorporated error-handling mechanisms to manage potential issues, such as connection failures when URLs were nonexistent. The code is explained in detail in Multimedia Appendix 2 and is publicly accessible on GitHub [41].

The web crawler requested 50,000 URLs from websites linked to the Assets and Health platform, comprising 25,000 activities and 25,000 resources, to capture reference numbers from the onset of the program, from July 2015 to December 23, 2022, which was the day when the data were scraped. The program "parser" then analyzed the HTML code of each existing URL and extracted the desired elements, stripping the text, which was automatically stored in 2 CSV data sets for activities and resources. The encoding of the data sets was revised to avoid mismatches between the character set used to represent the text data and that of the scraped text, as this can result in certain characters being displayed as symbols. An initial review of scrapped health assets was conducted to exclude irrelevant observations. We filtered out assets registered outside Catalonia or targeted solely at children and youth before proceeding with the text processing and mining steps.

Steps 4 to 6: Text Processing and Mining to Generate Meaningful Contextual Variables

Step 4: Cleaning and Preprocessing

Cluttered and inconsistent text data obtained from web scraping were preprocessed for analysis [31]. We used RStudio (version 2022.12.0), with the "tm" [42] and "qdap" [43] libraries. The "tm" library provides functions for cleaning, preprocessing, and analyzing text data. The "qdap" library allows text categorization, word frequency calculation, tokenization, and clustering.

The first author manually examined a set of activities and resources from scraped text to assess the quality and structure of the text data. This step was crucial for identifying inconsistencies, such as assigning different age ranges (eg, 60 and 65 years) to older adults simultaneously.

To preprocess the text data, the columns containing free text, namely titles and descriptions of the health assets, were merged and converted to "corpus" objects—a data structure for text data in R. Columns with text derived from fixed responses were not preprocessed. Irrelevant stop words were then removed, text data were segmented into individual units that could be transformed into numerical variables (tokenization), and words were normalized to their root form (stemming) [44]. The 2 data sets, containing health assets registered as activities and resources, were processed independently. The code with explanations for this step can be accessed through GitHub [45].

Step 5: Defining Dictionaries and Categorizing Text

We used text mining techniques to develop a classification system to filter and categorize health assets for older adults to enhance social connections from all other registered activities and resources on the platform. Text classification is pivotal for the generation of new variables of interest from unstructured data, because it can categorize text into predefined classes or labels.

First, to classify health assets, we predefined new variables and categories created through a deductive approach, based on the literature on social connections [46,47]. We also used inductive processes to create new variables and categories based on patterns identified during the text analysis and discussion among the research team. Table 1 lists the new variables, categories, type of creation process, and literature sources. Detailed definitions of the new variables and categories are provided in Multimedia Appendix 3 [46,48-51].

Second, we created document-term matrixes from the corpus of preprocessed text data containing health asset titles and descriptions. A document-term matrix is a mathematical matrix that describes the frequency of terms in a collection of textual data [31]. The frequency of each word was calculated and sorted based on their frequency values, representing the number of times a term appeared in the title and description of health assets.

Third, over the course of 3 meetings, 2 researchers (PG-H and CM) identified and selected high-frequency words that were repeated 15 times or more in the scraped data, grouped them into topic-specific dictionaries, and refined the list. Eligibility criteria, informed by the definitions of each new variable category, were developed to determine which words to include in each dictionary.

Finally, a classification system was developed using a rule-based classifier. A rule-based classifier categorizes data into predefined classes by applying a set of human-defined rules and conditions based on the features and attributes of the data [52]. We opted for a rule-based system over more complex machine learning classifiers, as this approach is better suited for scenarios with a limited number of specific labels and smaller data sets and ensures efficiency and interpretability without the need for extensive training data [52,53]. Topic-specific dictionaries consist of lists of words related to the definitions of the predefined variable categories as conditions to classify the text data [54]. Finally, an R function was developed to automatically generate a new column for each new variable, search for dictionary words in the scraped data, and assign a new category value if a word was found. The classifier system, including topic-specific dictionaries, is accessible on GitHub [45].

Table 1. New variables and categories created for the classification system of health assets.

New variables	Categories within each variable	Source columns from scraped text data	Creation process and literature sources
Activities			
Activity type	Leisure and skill development, physical activity, social facilitation, psychological therapies, awareness campaigns, health and social care, and befriending	Title and description	Deductive [46,48,49]
Format	Group and individual	Title and description	Deductive [49]
Focus	Direct and indirect	Title and description	Deductive [49]
Age	Children, youth, adults, older adults, general population, minors unspecified, and adults unspecified	Description, target population, and topics	Deductive [50]
Gender	Women, men, nonbinary, and any	Description, target population, and topics	Deductive [51]
Vulnerable populations	Migrants, caregivers, substance use, physical diseases, risk social exclusion, mental diseases, and all[a]	Title, description, target population, and activity topics	Inductive
Resources			
Resource type	Municipal natural and green space, health institution, social welfare institution, education institution, patient advocacy group, charitable and voluntary organization, faith-based organization, parent school associations, public library, civic center, sports institution, leisure and cultural association, neighborhood association, and cultural institution	Title and description	Inductive
Focus	Direct and indirect	Title and description	Deductive [49]
Age	Children, youth, adults, older adults, general population, minors unspecified, and adults unspecified	Title, description, and topics	Deductive [50]
Gender	Women, men, nonbinary, and any	Description and topics	Deductive [51]
Vulnerable populations	Migrants, caregivers, substance use, physical diseases, risk social exclusion, mental diseases, and all[a]	Title, description, and topics	Inductive

[a] "Substance use," "physical diseases," "risk social exclusion," "mental diseases," and "all" are simplified terms for target populations experiencing substance use, physical diseases, mental diseases, those at risk of social exclusion, and the general population.

Step 6: Manual Verification

The categories assigned to the new variables for each health asset were reviewed for inconsistencies by 2 researchers with expertise in the topic (PG-H and Angeli Chacaliaza). Manual verification refers to a one-by-one examination of the classified data by human reviewers to ensure the accuracy of the new variables created [55]. Verification involved an independent review of 200 health assets by 2 researchers to assess new variable categories based on eligibility criteria. Discrepancies were resolved through web meetings with manual reclassification if necessary. The data sets were then divided into groups of 500 health assets for independent review. Agreement rates between the verified and automatically generated variables were computed using the Excel software.

Step 7 and 8: Spatial Overlay Analysis to Locate Observations

Step 7: Identifying the Layers of Spatial Data

We used a spatial overlay analysis to generate a new variable that identified the BHA in which each health asset was located. The analysis was conducted in RStudio (version 2022.12.0) because of its many packages specifically designed for spatial overlay analysis, such as "sp," [56] "sf," [57] "rgdal," [58] "rgeos," [59] and "ggplot2" [60].

In this step, 2 spatial layers were identified. The first layer consisted of polygonal data depicting 374 BHAs in Catalonia. The data were obtained from the open database of the General Directorate of Health Planning and Research in Catalonia. Polygonal data can represent geographic or jurisdictional regions by defining their boundaries [61]. The second layer comprises a vector of geographic point data for each health asset. Point data consisted of longitude and latitude coordinates obtained from the addresses scraped for each activity and resource using Excel add-on GeoCode, a map tool that uses Google services to automatically retrieve longitudes and latitudes from addresses.

Step 8: Matching Coordinate Reference System and Finding Intersection Points

Step 8 involves transforming the spatial data layers into a common coordinate reference system (CRS) and identifying the intersecting points. The CRS of a spatial object determines its location on the Earth's surface. Thus, analyzing 2 or more spatial layers with different CRS can produce misleading outcomes [61]. To identify areas of overlap between health asset coordinates and BHAs, the following steps were taken: (1) coordinates were transformed to a simple feature object format, (2) simple feature objects were converted into single points using the "st_point" function, and (3) both spatial data layers were converted to a common CRS.

Finally, a spatial overlay analysis was performed using the “st_intersects” function to determine the BHA polygons with which each health asset point data intersected. The function was applied in a loop to each row of the activity and resource data sets. The resulting outputs are stored in new columns named “Code_BHA” and “Name_BHA.” The code, along with explanations for steps 7 and 8, is available in GitHub [45].

Data Set Filtering and Descriptive Analysis

The new data sets were filtered using the new variables and categories to select health assets with the potential to foster social connections among older adults from all scrapped assets. Eligible health assets registered as activities and resources were included if (1) the target population included older adults, (2) the format was either group activities or individual activities fostering social connections (eg, befriending), and (3) they were located in Catalonia.

A descriptive analysis was conducted in RStudio (version 2022.12.0), to understand the characteristics and asset registration trends of stakeholders across BHAs. Frequencies and proportions were calculated for each category of the new variables (activity type, format, focus, age, sex, and vulnerable populations). Temporal registration trends of activities and resources were analyzed using time-series graphs with local polynomial regression fitting lines, a nonparametric method used to describe the deterministic variation in data [62]. We also computed the average weekly registration of activities and resources in each BHA, assuming a Poisson distribution, where λ represents the weekly health asset registrations per area. Finally, visualization techniques were used to analyze the temporal evolution of the registration of activities and resources on the Assets and Health websites across BHAs, as well as their geographic distribution.

Ethical Considerations

The data collected in this study were publicly accessible and did not contain any personal or sensitive information. Thus, ethical approval and participant consent were not required for this study. In addition, before data collection, we verified that the websites of interest did not have any explicit prohibitions against automatic web scraping, such as a “robots.txt” file or similar declarations.

Results

Results From WeTMS

Web Scraping

Of the 50,000 URLs inspected, 17,305 contained websites describing health assets (9558 activities and 7747 resources) registered with local stakeholders from July 2015 to December 2022. The number of observations obtained through web scraping matched the total number of assets reported on the Assets and Health platform, thus demonstrating the efficacy of the web scraper. No missing values were detected for the main variables (eg, title, description, location, and date of asset registration). In the activity data set, 9.56% (480/5022) of observations did not disclose the *activity cost*, and 49.04% (2463/5022) did not report the *activity duration*. An example of an activity and resource, as they appear in the scraped data sets, is provided in Textbox 1.

Textbox 1. Example of a health asset registered as activity and resource extracted from the scraped text (English translation).

Activity row #860

- Title: School for Adults
- Description: Reading and writing classes
- Population: Over 65 years old—anyone (district neighbors over 65 years old)
- Location: Campoamor Street 92, 08204, Civic Center Rogelio Soto, Sabadell, Barcelona, Catalonia, Spain
- Organizations: Civic Center Rogelio Soto, Campoamor Neighborhood Association
- Registration date: February 6, 2020
- Is free: Yes
- Categories: Women, older adults, people at risk of exclusion, school of health, mental health, or emotional well-being
- Time activity: From September 13, 2019, to June 30, 2020

Resource row #1961

- Title: Association of Retirees and Pensioners, La Pineda
- Description: Association that aims to promote cultural training and sports activities for older adults, as well as avoiding loneliness and social isolation, fostering relationships between them
- Registration date: May 15, 2017
- Location: Alfredo Kraus Street 20, 43481, La Pineda Vila-seca, Tarragona, Catalonia, Spain
- Categories: Older adults, mental health or emotional well-being, physical activity, community health

Text Mining

From the text processing of the corpus of titles and descriptions, a total of 12,560 tokens (or raw words) were identified for activities and 7301 for resources, of which 996 (7.9%) and 594 (8.1%) words had a frequency >15. Using words with a frequency of >15, we constructed 73 topic-specific dictionaries corresponding to each category of the new variables. For instance, for the *physical activity* category under the *activity type* variable, the topic-specific dictionary included words such as “physical,” “exercise,” “gym,” “yoga,” and “sport.” Figure 2 presents popular dictionary words for each category within the *activity type* variable.

After applying the rule-based classifier using topic-specific dictionaries, manual verification of the output yielded variable levels of agreement ranging from 62.02% (3417/5509) to 99.47% (4886/4912) across variables. For instance, variables with lower classification accuracy had a larger number of possible categories, a more evenly distributed number of observations across categories, or words repeated fewer than 15 times within the title and description corpus. The agreement rates between the verified and automatically generated databases are presented in Table 2.

Figure 2. Categories within the “activity type” variable showcasing popular words derived from topic-specific dictionaries (English translation).

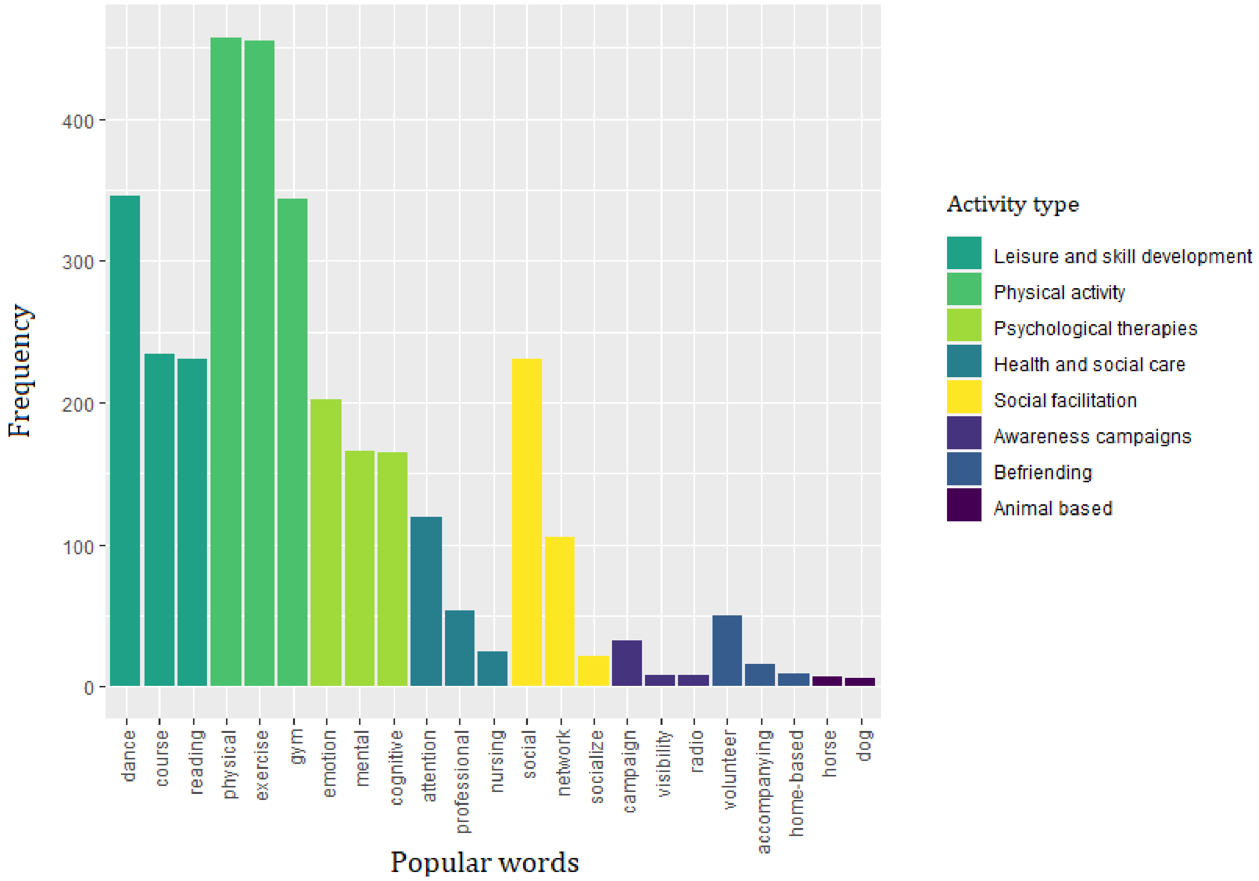


Table 2. Agreement rate between manually verified and automatically classified data sets.

New variables generated	Correctly assigned categories, n (%)
Activities data set (n=6260)	
Age	4855 (77.55)
Gender	6215 (99.28)
Vulnerable populations	5525 (88.26)
Activity type[a]	3417 (62.02)
Format[a]	5326 (96.67)
Focus[a]	5342 (96.97)
Resources data set (n=4912)	
Age	4029 (82.02)
Gender	4843 (98.59)
Vulnerable populations	3883 (79.05)
Resource type	3845 (78.28)
Focus	4886 (99.47)

[a] Automatic classification of the categories for variables *activity type*, *format*, and *focus* was performed only for activities targeting older adults (n=5509).

Spatial Overlay Analysis

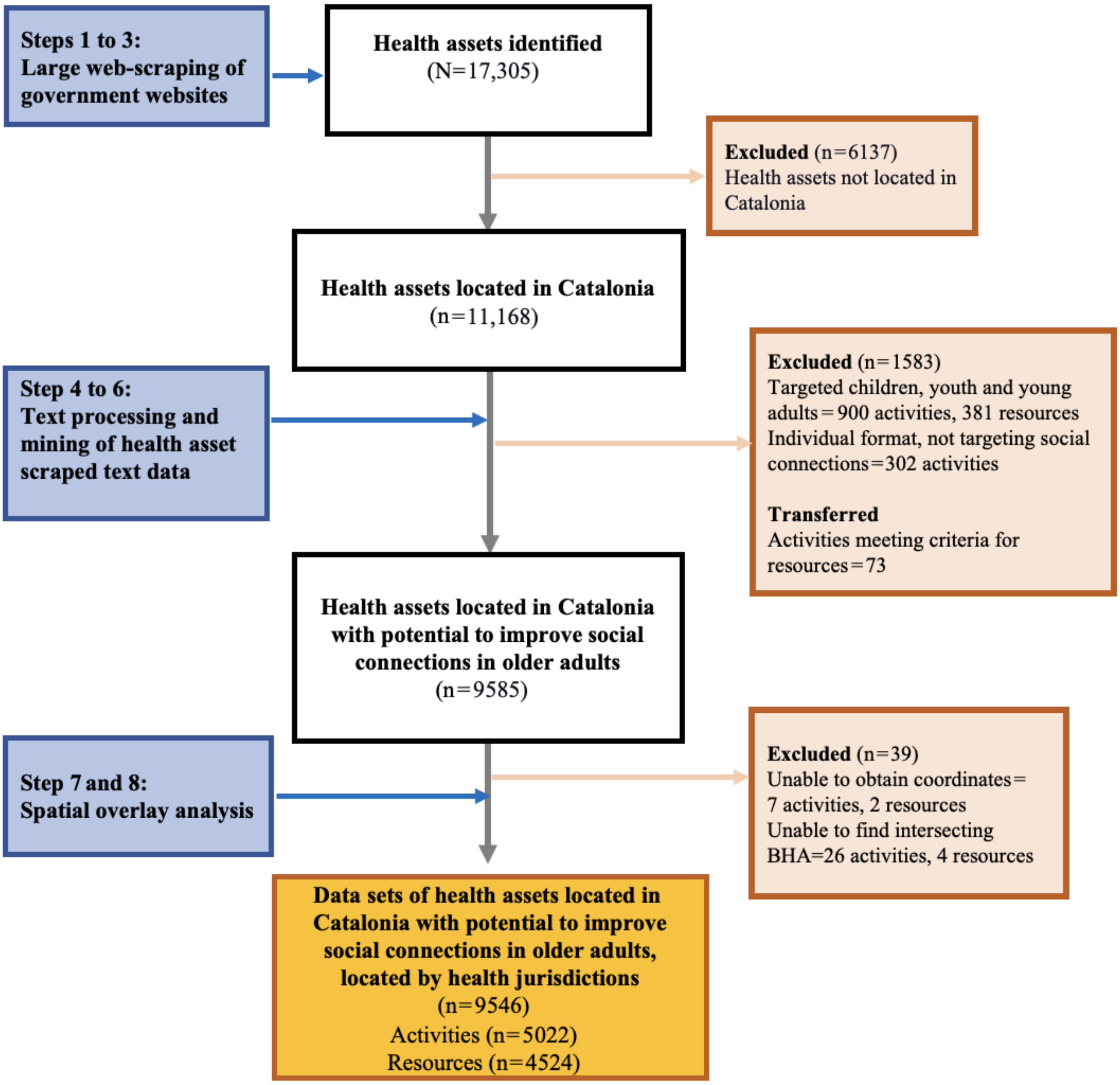
Coordinates for the locations of 0.36% (18/5055) of activities and 2.41% (109/4530) of resources were not identified. Manual searches on Google Maps using addresses allowed us to locate the coordinates for all but 7 activities and 2 resources. Through spatial overlay analysis, intersections between spatial points and BHAs were not identified for 26 activities or 4 resources. The newly generated columns for the variables *Code_BHA* and *Name_BHA* encompassed values for 318 distinct BHAs, representing 85% of the 374 BHAs.

Results From Data Set Filtering and Descriptive Analysis

Filtering Health Assets With Potential to Enhance Older Adults' Social Connections

Using the newly generated contextual variables, we filtered the data sets of activities and resources to identify those with the potential to foster social connections among older adults. From the initial 17,305 health assets identified, we obtained 9546 eligible health assets, comprising 5022 activities and 4524 resources. The reasons for exclusion and the stages in which health assets were discarded are shown in Figure 3.

Figure 3. Flowchart for the filtering of health asset data sets by contextual variables related to social connection constructs generated using the web scraping, text mining, and spatial overlay (WeTMS) method. BHA: basic health area.


Steps 1 to 3:
Large web-scraping of government websites
Health assets identified
(N=17,305)
Excluded (n=6137)
Health assets not located in Catalonia
Health assets located in Catalonia
(n=11,168)
Step 4 to 6:
Text processing and mining of health asset scraped text data
Excluded (n=1583)
Targeted children, youth and young adults=900 activities, 381 resources
Individual format, not targeting social connections=302 activities
Transferred
Activities meeting criteria for resources=73
Health assets located in Catalonia with potential to improve social connections in older adults
(n=9585)
Step 7 and 8:
Spatial overlay analysis
Excluded (n=39)
Unable to obtain coordinates= 7 activities, 2 resources
Unable to find intersecting BHA=26 activities, 4 resources
Data sets of health assets located in Catalonia with potential to improve social connections in older adults, located by health jurisdictions
(n=9546)
Activities (n=5022)
Resources (n=4524)

Characteristics of Eligible Health Assets

Of the health assets registered as activities, 24.59% (1235/5022) specifically targeted older adults, whereas 75.41% (3787/5022) targeted broader age ranges, including the older population. Most resources targeted the general population and included older adults, with only 4.12% (207/5022) being exclusively for older adults, such as civic centers for retirees. Only 2.49% (238/9546) of the health assets had a sex-specific target; these were predominantly women (n=212). Among all health assets, 13.5% (678/5022) of activities and 7.98% (361/4524) of resources were tailored for specific vulnerable groups, with physical or mental illness being the primary focus.

Group-oriented activities promoting social interactions accounted for 99.56% (5000/5022) of the eligible activities. However, only 4.36% (219/5022) explicitly used concepts related to social connections (eg, loneliness and social isolation) in titles and descriptions. Over 57% (2862/5022) of the activities were cost-free, and the most common activity duration was 1 to 3 months (975/5022, 19.41%). Data on format, duration, and cost are not available for resources.

The analysis of the new variable *activity type* showed that leisure and skill development activities were most common (1844/5022, 36.72%). This included group handcrafts, dance, painting, theater, cooking, choir courses, and conversation groups that focused on shared-interest topics. Group exercise activities (eg, walking groups) accounted for 31.08% (1561/5022) of the activities. Over 22% (1103/5022) of the activities involved group activities with health and social professionals outside the health care center, including psychological therapies and health and social care. Finally, 8.46% (425/5022) were social facilitation activities such as group meetings to share common interests (eg, film forums). Overall, more than half of these activities were registered between 2021 and 2022 (Figure 4).

Almost 61.49% (2782/4524) of the registered resources facilitated exchange of knowledge and interests among older adults. These resources included leisure and cultural associations; public libraries; civic centers; and cultural, sports, and educational institutions. Municipal natural and green spaces where adults can gather accounted for 17.28% (782/4524) of the resources. A total of 595 (13.1%) health institutions and 140 (3.1%) social welfare institutions were found, including primary care centers, health and social foundations, and advocacy institutions promoting social inclusion. Other resources linked to health and social welfare include patient advocacy groups, faith-based organizations, and charitable and voluntary organizations. In contrast, most resources were registered before 2019 (Figure 5). A detailed descriptive analysis of the type, target population, focus, cost, format, and duration of health assets is included in Multimedia Appendix 4.

Figure 4. Number of activities with the potential to enhance older adults' social connections by type and year of registration (2015-2022).

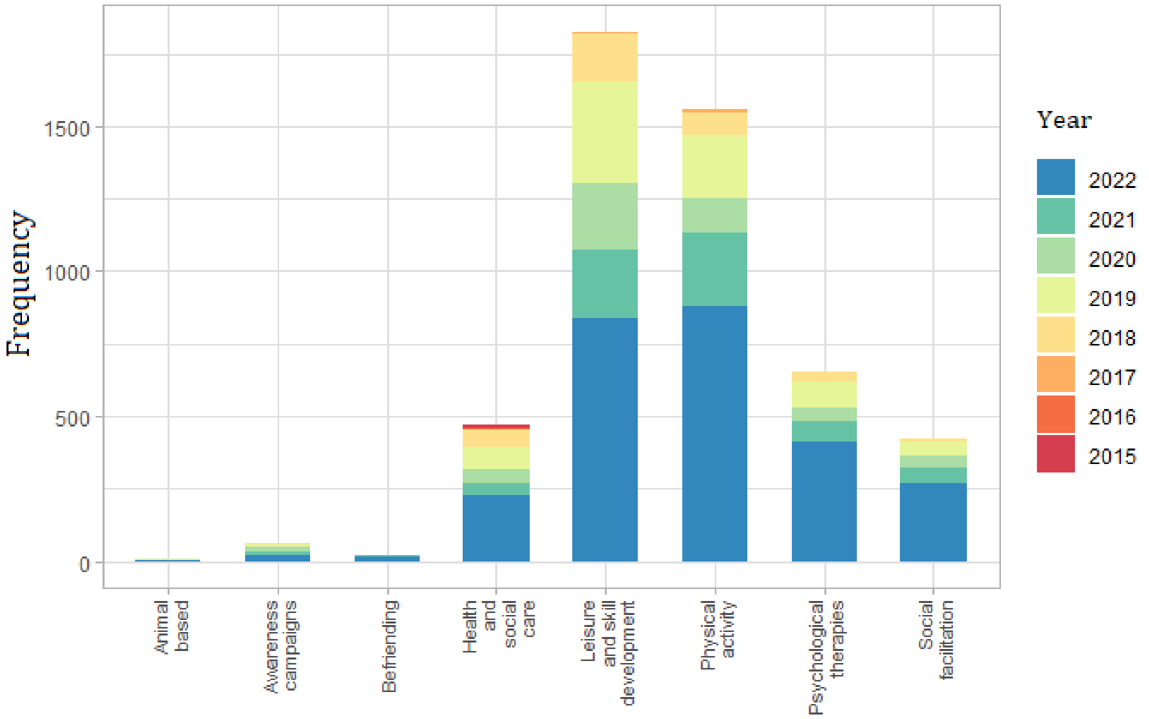
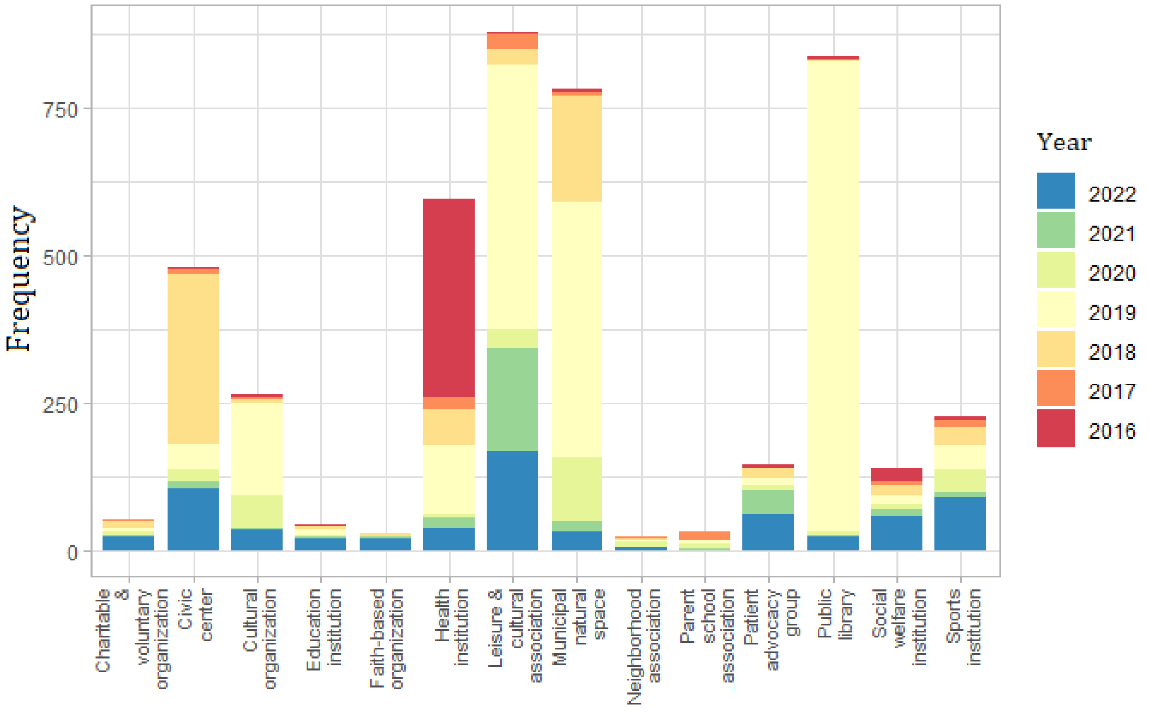


Figure 5. Number of resources with potential to enhance older adults' social connections, by type and year of registration (2015-2022).



Overview of Registration Trends of Eligible Health Assets Across BHAs

The first registry of a health asset on the Assets and Health websites occurred on July 23, 2015, and the last on December 23, 2022, the day when web scraping was conducted. Total registration of activities remained consistently low from the start of the program until early 2018, whereas for resources, a registration peak was observed in late 2016. Activity and resource registrations have increased from 2018 to mid-2020. A decline in registration was observed from early 2020 to mid-2021, coinciding with the outbreak of the COVID-19 pandemic. Local polynomial regression fitting lines showed a growing pattern in the registration of activities from 2021 onwards, whereas resource registration remained low (Figure 6).

On the basis of the observed trends, 4 implementation periods were defined to better understand registration trends: period 1, from July 2015 to January 2018; period 2, from February 2018 to February 2020; period 3, from March 2020 to May 2021; and period 4, from the end of June 2021 to December 2022. During the first two and a half years of the program, the average number of activities and resources registered per week across all BHAs was 0.37 and 3.47, respectively, increasing to 11.83 and 25.27 in period 2. During the COVID-19 pandemic in period 3, these figures decreased to an average of 3.19 activities and 20.64 resources per week. period 4 had the highest registration rate for activities (38.52/wk).

To calculate the registration trends in individual BHAs, we divided the number of BHAs with one or more activities registered by the total number of BHAs (n=374) for each period. We did not consider the resource data set because of the observed patterns suggesting centralized registration, rather than local registration. For instance, in late 2016, resources in 237 BHAs were registered in a single day. At the end of period 1, 8% (30/374) of the BHAs had one or more registered activities, which increased to 85% (318/374) by the end of period 4 (Figure 7; Table 3). The number of health assets registered per BHA varied significantly, ranging from 0 to 263 activities and 0 to 265 resources. The median number of activities registered per BHA from 2015 to 2022 was 5 (IQR 13.75) and 9 (IQR 10) for resources. Figure 8 illustrates the geographic distribution of the activities and resources registered in each period.

Figure 6. Weekly registration trends of health assets from July 2015 to December 2022. The y-axis represents the number of registered activities and resources per week. The x-axis represents time, labeled by years for clarity.

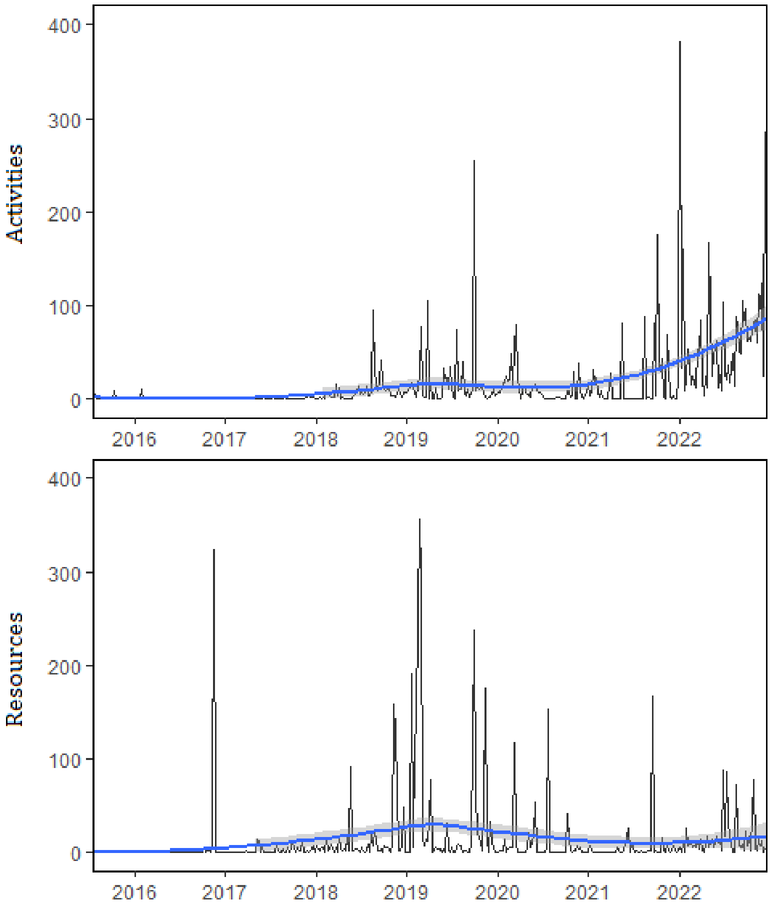


Figure 7. Cumulative frequency of basic health areas (BHAs) with registered health assets aimed at enhancing older adults' social connections from July 2015 to December 2022. Each point represents a BHA at the time of its first asset registration on the Assets and Health platform, cumulative.

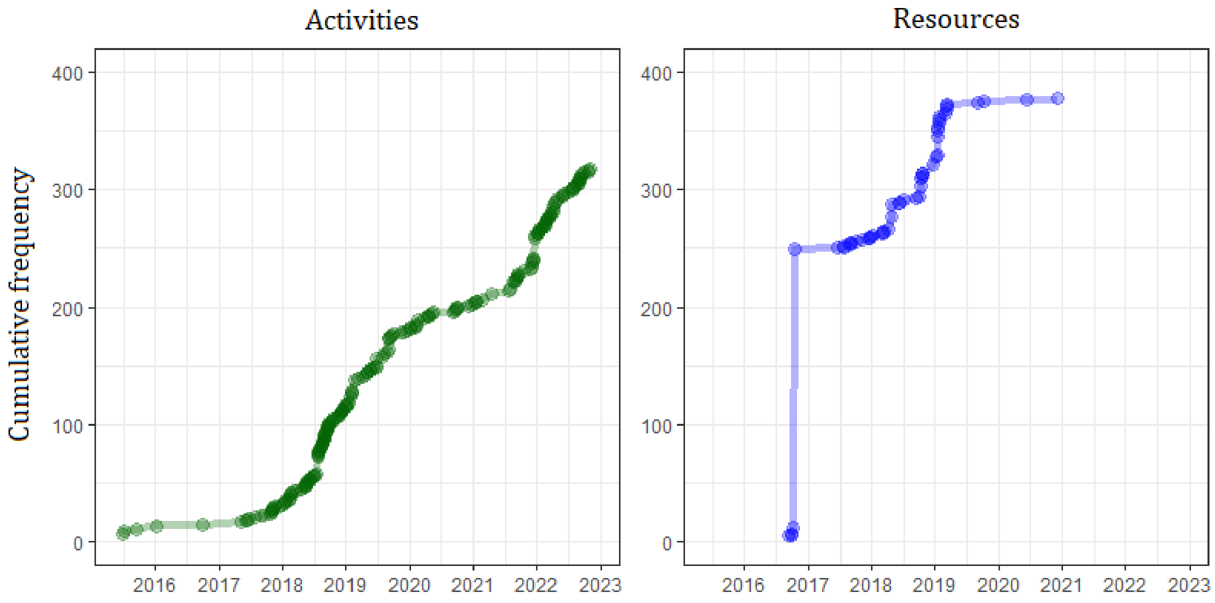


Table 3. Average number of health assets registered per week and proportion of basic health areas (BHAs) with one or more activities registered per period.

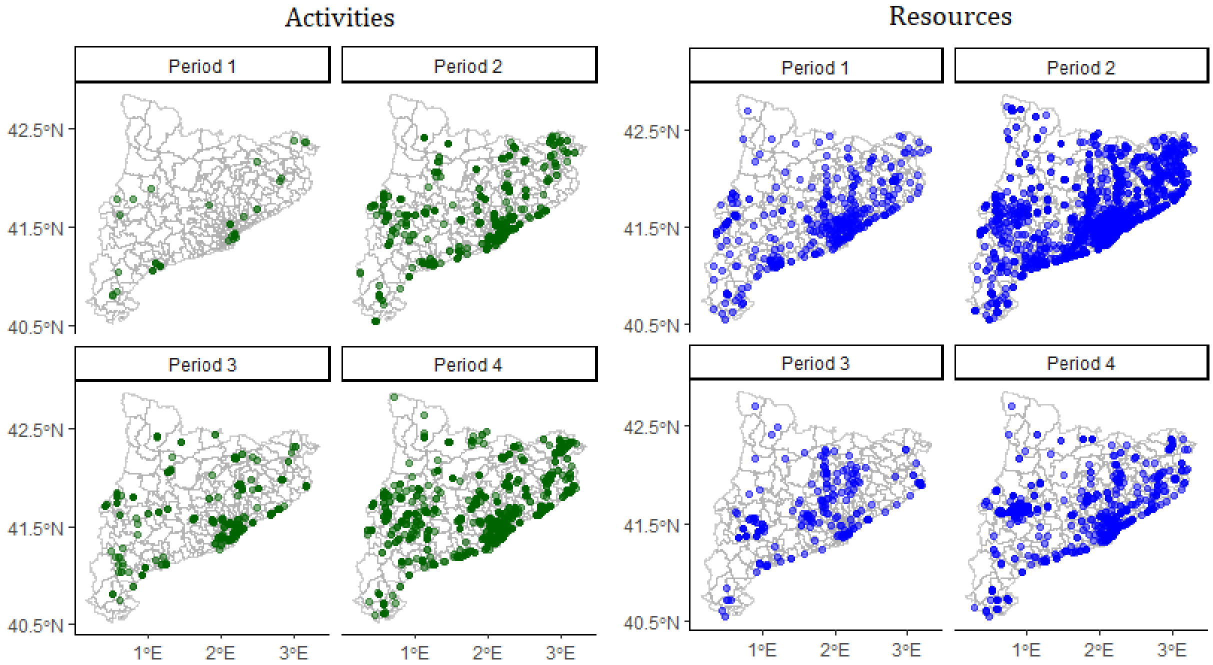
Time periods	Number of health assets		λ (95% CI)[a]		BHAs with registered activities	
	Activities	Resources	Activities	Resources	Values, n[b]	Cumulative proportion of BHA (%)[c]
Period 1: July 2015 to January 2018	50	462	0.37 (0.27-0.48)	3.47 (3.16-3.79)	30	8
Period 2: February 2018 to February 2020	1290	2755	11.83 (11.19-12.48)	25.27 (24.33-26.21)	154	49.2
Period 3: March 2020 to May 2021	562	301	8.64 (7.93-9.36)	4.63 (4.11-5.15)	28	56.6
Period 4: June 2021 to December 2022	3120	941	38.52 (37.16-39.87)	11.62 (10.87-12.36)	106	85

[a] λ denotes the average number of health assets registered per week in each period.

[b] Unique BHAs that registered activities targeting social connections in older adults for the first time in the specified period.

[c] BHAs that registered such activities up to and including each period with earlier registrations.

Figure 8. Geographic distribution of activities and resources with potential to enhance older adults' social connections across basic health areas.



Discussion

Innovation: Generating Area-Specific Contextual Variables From Unstructured Web-Based Data

We introduce a novel approach for generating area-specific contextual variables from unstructured website data using WeTMS. By combining the methods commonly used in computer and data science, we were able to efficiently gather and transform large amounts of website data into comprehensive data sets of theoretically informed variables. The resulting data sets enabled us to identify and characterize health assets with the potential to enhance social connections among older adults registered within health jurisdictional areas from 2015 to 2022. In addition, this approach allowed us to examine area-specific registration trends for health assets, showing the use of the Assets and Health platform developed as part of a public health strategy in Catalonia. We provided detailed explanations of concepts, steps, and the code used, and included supplementary information to facilitate the replication of the steps, attempting to familiarize novice readers with these techniques.

Applications of the WeTMS Method

Our method provides a tool for researchers interested in developing new contextual variables when data are scarce or difficult to obtain using traditional means. Researchers in fields such as public health, nursing, and social epidemiology who study the impact of emerging health and social phenomena on health outcomes and determinants of health can benefit from this method. A practical example of an emerging social determinant of health, such as precarious employment [63], can consist of applying the web-scraping steps to obtain website data from employment portals, text mining to analyze posts, identifying precarious job offers, and spatial overlay analysis to locate them into geographic areas and study the effect on population outcomes using multilevel modeling. Researchers and program evaluators in health services research and implementation science can use this method to obtain data to conduct descriptive analyses explaining policy adoption within jurisdictional or geographic areas, following the research case outlined in this study.

A key feature of this method is that its steps can be implemented in sequence or independently, depending on the research goals. For example, researchers interested in generating new variables from text data without locating them in specific geographic areas can follow steps 1 to 6, which involve web scraping, text processing, and mining. If a data set is already available and researchers want to group the data by geographic settings, they can follow steps 7 and 8, which involve overlay spatial analysis.

Challenges and Limitations

There are challenges with this method that can limit its feasibility and application. In our example, extracting comparable data from multiple URLs was feasible because the websites associated with the Assets and Health platforms had similar HTML structures. The consistent placement of targeted information across websites simplifies the complexity of the web-scraping program. Thus, the attached web scraper code is only suitable for single or multiple websites with a limited number of distinct HTML structures (eg, forums, social media, and employment portals). Studies in which the target data are spread over different websites with varied designs require more advanced programming [11].

A key step in the process—the creation of topic-specific dictionaries for the classification of observations—necessitates a deep understanding of the field and the terminology used in the data. Overall, the rule-based classifier demonstrated high accuracy. However, some variables, such as *activity type*, showed a higher rate of errors, in part because the dictionaries used to classify them contained only high-frequency words found in the titles and descriptions. Thus, our experience suggests that manual verification of new variables and categories by researchers with a comprehensive understanding of the data and subject matter is essential to ensure data validity before statistical analysis. However, this can be unfeasible for large data sets. In such scenarios, the impracticality of manual verification may necessitate the use of complex machine learning classifiers, presenting a trade-off in the confidence of the data that potentially compromises the robustness of the resulting variables [52].

Spatial overlay analysis effectively localizes health assets to their respective health jurisdictions, facilitated by the acquisition of complete addresses during the web scraping phase and the availability of a high-quality polygon map for analysis. Geospatial maps can be obtained from government agencies, nonprofit organizations, and commercial providers. If maps are unavailable, they can be created using accessible satellite imagery [64]. However, the necessity for location-specific data (eg, addresses, postal codes, and cities) for each observation to generate contextual-level variables limits the range of suitable data sources available to researchers.

Ethical and data protection considerations are important. Web scraping is typically permitted when data are publicly accessible and not subject to international legislation concerning personal data, trademarks, copyrights, or private information [30]. Automatic extraction of internet data might be unfeasible if the data are not publicly available or if a website's terms of service restrict automated collection and analysis [65]. Researchers may consult ethics bodies to ensure that the methodology adheres to ethical standards when dealing with sensitive topics and personal information, even when relying on publicly available sources.

In addition to these challenges, the method and data sets that it produces have limitations. The complexity of these steps requires introductory technical knowledge. Thus, we have provided detailed explanations and supplementary information that can support researchers, as they familiarize themselves with the steps. We anticipate that the compendium of concepts, code, software packages, and references gathered from trustworthy sources will serve as a resource for those interested in these techniques.

Another limitation is the bias associated with the classifier system. The development of classification systems inherently relies on the subjective judgment of researchers. This can result in misclassifications, particularly those related to assumptions about race, gender, or social exclusion factors, especially within machine learning classifiers [66]. It is advisable for researchers to engage in a reflexive process, carefully considering their assumptions in the definition and selection of dictionary words and to critically evaluate how these decisions may influence the investigation [67].

Finally, although the data sets generated are robust for descriptive analysis, researchers should proceed with clearly defined assumptions when using new context-level variables in statistical analyses, particularly in multilevel modeling or ecologic studies that aim to draw inferences. Although we successfully compiled 2 data sets of health assets from targeted websites, their comprehensiveness and accuracy in reflecting all identified assets across BHAs remain unknown. It is possible that some local organizations in BHAs were more or less likely to register assets on Assets and Health websites, influenced by context-specific factors such as management support and training on the platform [68]. If feasible and ethical, it would be advisable for researchers to triangulate data to validate the data sets, thus verifying web-scraped data with secondary sources or direct inputs from stakeholders [69].

Conclusions

The sequential use of WeTMS enabled the efficient creation of data sets of health assets registered with the Assets and Health websites in Catalonia, Spain, which aimed to enhance the social connections of older adults in local health jurisdictions. Our descriptive analysis demonstrated the usefulness of the data sets in exploring the characteristics of contextual variables, as well as in understanding temporal patterns and spatial distributions.

Contextual-level variables generated via WeTMS may also be used in hierarchical analyses to evaluate the impact of contextual factors on health outcomes when more robust sources, such as census data, are not available. Adherence to data protection standards and ethical considerations should also guide this process. Although WeTMS has potential value for multiple research disciplines, it presents challenges and limitations, including the need for internet data sources to have comparable structures, a dependence on location data, the potential lack of representativeness in website content, the requirement for technical expertise, and a significant time investment for manual verification.

Acknowledgments

The authors thank Angeli Chacaliaza for his valuable contribution as a second reviewer in the manual verification stage and Ann Tourangeau, Martine Puts, and Tina Behrouzi for their comments and suggestions on the first draft. The authors also acknowledge Charles Anderson for his technical assistance. This study was not supported by any funding. The authors attest that there was no use of generative artificial intelligence technology in the generation of text, figures, or other informational content in this manuscript.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

PG-H and CM conceptualized the study; PG-H wrote the study protocol, developed the method, and drafted the manuscript; and AG-V, LG-P, and CM provided expert input and conducted manuscript review and editing. All the authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional details in step 2: identifying hidden URLs, finding HTML elements, and further references.
[DOCX File , 733 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed explanation of Python libraries and web scraper code.
[DOCX File , 18 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Definitions of new variables and categories for text classification.
[DOCX File , 27 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Extended descriptive analysis of the type, target population, focus, cost, format, and duration of health assets.
[DOCX File , 21 KB-Multimedia Appendix 4]

References

1. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health. Feb 1998;88(2):216-222. [doi: 10.2105/ajph.88.2.216] [Medline: 9491010]
2. Sampson RJ, Raudenbush SW, Earls F. Neighborhoods and violent crime: a multilevel study of collective efficacy. Science. Aug 15, 1997;277(5328):918-924. [doi: 10.1126/science.277.5328.918] [Medline: 9252316]
3. Subramanian SV, Lochner KA, Kawachi I. Neighborhood differences in social capital: a compositional artifact or a contextual construct? Health Place. Mar 2003;9(1):33-44. [FREE Full text] [doi: 10.1016/s1353-8292(02)00028-x] [Medline: 12609471]
4. van den Berg AE, Maas J, Verheij RA, Groenewegen PP. Green space as a buffer between stressful life events and health. Soc Sci Med. Apr 2010;70(8):1203-1210. [doi: 10.1016/j.socscimed.2010.01.002] [Medline: 20163905]
5. Hox JJ, Moerbeek M, van de Schoot R. Multilevel Analysis: Techniques and Applications, Third Edition. Milton Park, UK. Taylor & Francis; 2017.
6. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J Epidemiol Community Health. Dec 10, 2012;66(12):1182-1186. [FREE Full text] [doi: 10.1136/jech-2011-200375] [Medline: 22577181]
7. Inter-ministerial Public Health Plan (PINSAP). Agencia de Salut Publica de Catalunya. Feb 14, 2014. URL: https://salutpublica.gencat.cat/web/.content/minisite/aspcat/sobre_lagencia/pinsap/continguts_antics/pinsap-en.pdf [accessed 2022-12-10]
8. Crane M, Bohn-Goldbaum E, Grunseit A, Bauman A. Using natural experiments to improve public health evidence: a review of context and utility for obesity prevention. Health Res Policy Syst. May 18, 2020;18(1):48. [FREE Full text] [doi: 10.1186/s12961-020-00564-2] [Medline: 32423438]
9. Leatherdale ST. Natural experiment methodology for research: a review of how different methods can support real-world research. Int J Soc Res Methodol. Jul 02, 2018;22(1):19-35. [doi: 10.1080/13645579.2018.1488449]
10. Diouf R, Sarr EN, Sall O, Birregah B, Bousso M, Mbaye SN. Web scraping: state-of-the-art and areas of application. In: Proceedings of the IEEE International Conference on Big Data (Big Data). Presented at: IEEE International Conference on Big Data (Big Data); December 09-12, 2019, 2019; Los Angeles, CA. URL: https://ieeexplore.ieee.org/document/9005594/authors#authors [doi: 10.1109/bigdata47090.2019.9005594]
11. vanden Broucke S, Baesens B. Practical Web Scraping for Data Science: Best Practices and Examples with Python. New York, NY. Apress; 2018.
12. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res. Mar 27, 2009;11(1):e11. [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]

13. Mackey TK, Li J, Purushothaman V, Nali M, Shah N, Bardier C, et al. Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: infoveillance study on Twitter and Instagram. JMIR Public Health Surveill. Aug 25, 2020;6(3):e20794. [FREE Full text] [doi: 10.2196/20794] [Medline: 32750006]
14. Li J, Xu Q, Shah N, Mackey TK. A machine learning approach for the detection and characterization of illicit drug dealers on Instagram: model evaluation study. J Med Internet Res. Jun 15, 2019;21(6):e13803. [FREE Full text] [doi: 10.2196/13803] [Medline: 31199298]
15. Smith E, Michalski S, Knauth KH, Kaspar K, Reiter N, Peters J. Large-scale web scraping for problem gambling research: a case study of COVID-19 lockdown effects in Germany. J Gambl Stud. Sep 27, 2023;39(3):1487-1504. [FREE Full text] [doi: 10.1007/s10899-023-10187-1] [Medline: 36707481]
16. Gregory AL, Piff PK. Finding uncommon ground: extremist online forum engagement predicts integrative complexity. PLoS One. Jan 19, 2021;16(1):e0245651. [FREE Full text] [doi: 10.1371/journal.pone.0245651] [Medline: 33465152]
17. Kogan NE, Bolon I, Ray N, Alcoba G, Fernandez-Marquez JL, Müller MM, et al. Wet markets and food safety: TripAdvisor for improved global digital surveillance. JMIR Public Health Surveill. Apr 01, 2019;5(2):e11477. [FREE Full text] [doi: 10.2196/11477] [Medline: 30932867]
18. de Oliveira DV, Albuquerque UP. Cultural evolution and digital media: diffusion of fake news about COVID-19 on Twitter. SN Comput Sci. Aug 28, 2021;2(6):430. [FREE Full text] [doi: 10.1007/s42979-021-00836-w] [Medline: 34485922]
19. Gaikwad SV, Chaugule A, Patil P. Text mining methods and techniques. Int J Comput Appl. Jan 16, 2014;85(17):42-45. [doi: 10.5120/14937-3507]
20. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform. Jan 28, 2020;8(1):e16023. [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
21. Gruebner O, Sykora M, Lowe SR, Shankardass K, Trinquart L, Jackson T, et al. Mental health surveillance after the terrorist attacks in Paris. Lancet. May 2016;387(10034):2195-2196. [doi: 10.1016/s0140-6736(16)30602-x]
22. Boon-Itt S, Skunkan Y. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health Surveill. Nov 11, 2020;6(4):e21978. [FREE Full text] [doi: 10.2196/21978] [Medline: 33108310]
23. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PLoS One. Aug 1, 2014;9(8):e103408. [FREE Full text] [doi: 10.1371/journal.pone.0103408] [Medline: 25084530]
24. Chang KT. Introduction to Geographic Information Systems. Chicago, IL. McGraw-Hill Higher Education; 2006.
25. Shankardass K, Jerrett M, Milam J, Richardson J, Berhane K, McConnell R. Social environment and asthma: associations with crime and No Child Left Behind programmes. J Epidemiol Community Health. Oct 11, 2011;65(10):859-865. [FREE Full text] [doi: 10.1136/jech.2009.102806] [Medline: 21071562]
26. Hornby-Turner YC, Peel NM, Hubbard RE. Health assets in older age: a systematic review. BMJ Open. May 17, 2017;7(5):e013226. [FREE Full text] [doi: 10.1136/bmjopen-2016-013226] [Medline: 28515182]
27. Finder of assets and health. Public Health Agency of Catalonia (ASPCAT). URL: https://salutpublica.gencat.cat/ca/sobre_lagencia/Plans-estrategics/pinsap/Accions-eines-i-projectes-relacionats/actius-i-salut/cercador-dactius-i-salut/index.html#googtrans(ca|en) [accessed 2022-12-01]
28. Sáinz-Ruiz PA, Sanz-Valero J, Gea-Caballero V, Melo P, Nguyen TH, Suárez-Máximo JD, et al. Dimensions of community assets for health. A systematised review and meta-synthesis. Int J Environ Res Public Health. May 27, 2021;18(11):5758. [FREE Full text] [doi: 10.3390/ijerph18115758] [Medline: 34072002]
29. Oliver-Parra A, González-Viana A, Grupo de Trabajo de Indicadores Básicos de Salud por Área Básica (GT-IBS). [Facilitating community oriented primary health care. Basic health indicators by small areas in Catalonia]. Gac Sanit. 2020;34(2):204-207. [FREE Full text] [doi: 10.1016/j.gaceta.2019.05.012] [Medline: 31488325]
30. Mitchell R. Web Scraping with Python: Collecting More Data from the Modern Web. 2nd edition. Sebastopol, CA. O'Reilly Media; Apr 4, 2018.
31. Munzert S, Rubba C, Meißner P, Nyhuis D. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Hoboken, NJ. John Wiley & Sons; 2015.
32. HTML Elements. W3 Schools. 2023. URL: https://www.w3schools.com/html/html_elements.asp [accessed 2023-12-24]
33. HTML Tags. W3 Schools. 2023. URL: https://www.w3schools.com/tags/ [accessed 2023-12-25]
34. YouTube. Jan 23, 2021. URL: https://www.youtube.com/watch?v=XsL8JDkH-ec [accessed 2023-12-25]
35. Wu S. Web Scraping Basics. Towards Data Science. Jul 15, 2020. URL: https://towardsdatascience.com/web-scraping-basics-82f8b5acd45c [accessed 2023-12-25]
36. Shafer C. Python Tutorial: Web Scraping with BeautifulSoup and Requests. YouTube. Nov 18, 2017. URL: https://www.youtube.com/watch?v=ng2o98k983k [accessed 2023-12-25]
37. Python homepage. Python. URL: https://www.python.org/ [accessed 2023-12-18]
38. Reitz K. Requests documentation: release 2.31.0. Build Media. Aug 18, 2023. URL: https://buildmedia.readthedocs.org/media/pdf/requests/latest/requests.pdf [accessed 2023-12-18]
39. Richardson L. Beautiful Soup 4.12.0 documentation. Beautiful Soup. URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ [accessed 2022-12-10]
40. User guide - pandas 2.1.4 documentation. Pandas. URL: https://pandas.pydata.org/docs/user_guide/index.html#user-guide [accessed 2022-12-15]

41. Targeted web scraping implementation evaluation. GitHub. URL: https://github.com/paugalvez/Targeted_webscraping_implementation_evaluation.git [accessed 2023-12-18]
42. Feinerer I. Introduction to the tm Package Text Mining in R. The Comprehensive R Archive Network. Feb 5, 2023. URL: https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf [accessed 2023-12-18]
43. Rinker T, Goodrich B, Kurkiewicz D. qdap: bridging the gap between qualitative data and quantitative analysis. The Comprehensive R Archive Network. May 11, 2023. URL: https://CRAN.R-project.org/package=qdap [accessed 2023-12-18]
44. Vijayarani S, Janani R. Text mining: open source tokenization tools – an analysis. Adv Comput Intell Int J. Jan 30, 2016;3(1):37-47. [doi: 10.5121/acii.2016.3104]
45. Text analysis spatial overlay analysis R. GitHub. URL: https://github.com/paugalvez/Text_Analysis_Spatial_Overlay_Analysis_R [accessed 2023-12-18]
46. Gardiner C, Geldenhuys G, Gott M. Interventions to reduce social isolation and loneliness among older people: an integrative review. Health Soc Care Community. Mar 13, 2018;26(2):147-157. [FREE Full text] [doi: 10.1111/hsc.12367] [Medline: 27413007]
47. Galvez-Hernandez P, González-de Paz L, Muntaner C. Primary care-based interventions addressing social isolation and loneliness in older people: a scoping review. BMJ Open. Feb 04, 2022;12(2):e057729. [FREE Full text] [doi: 10.1136/bmjopen-2021-057729] [Medline: 35121608]
48. Freedman A, Nicolle J. Social isolation and loneliness: the new geriatric giants: approach for primary care. Can Fam Physician. Mar 2020;66(3):176-182. [FREE Full text] [Medline: 32165464]
49. National Academies of Sciences, Engineering, and Medicine; Division of Behavioral and Social Sciences and Education; Health and Medicine Division; Board on Behavioral, Cognitive, and Sensory Sciences; Board on Health Sciences Policy; Committee on the Health and Medical Dimensions of Social Isolation and Loneliness in Older Adults. Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System. Washington, DC. National Academies Press; 2020.
50. World population ageing 2015. United Nations, Department of Economic and Social Affairs, Population Division. 2015. URL: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf [accessed 2023-01-16]
51. Classification of gender. Statistics Canada. URL: https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1326727&CVD=1326727&CLV=0&MLV=1&D=1 [accessed 2023-01-16]
52. Vijayan VK, Bindu KR, Parameswaran L. A comprehensive study of text classification algorithms. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI). Presented at: International Conference on Advances in Computing, Communications and Informatics (ICACCI); September 13-16, 2017, 2017; Udupi, India. URL: https://ieeexplore.ieee.org/document/8125990 [doi: 10.1109/icacci.2017.8125990]
53. Radovanović M, Ivanović M. Text mining: approaches and applications. Novi Sad J Math. 2008;38(3):227-234. [FREE Full text]
54. Puschmann C, Haim M. Topic-specific dictionaries. Automated Content Analysis with R. URL: https://content-analysis-with-r.com/4-dictionaries.html [accessed 2023-01-04]
55. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296-298. [FREE Full text] [doi: 10.1197/jamia.M1733] [Medline: 15684123]
56. Pebesma EJ, Bivand RS. Classes and methods for spatial data in R. R News. Nov 2005;5(2):9-13. [FREE Full text]
57. Pebesma E. Simple features for R: standardized support for spatial vector data. R J. 2018;10(1):439-446. [doi: 10.32614/rj-2018-009]
58. Bivand R, Keitt T, Rowlingson B. rgdal: bindings for the 'geospatial' data abstraction library. The Comprehensive R Archive Network. Nov 21, 2017. URL: http://cran.nexr.com/web/packages/rgdal/index.html [accessed 2023-01-16]
59. Bivand R, Rundel C. rgeos: interface to geometry engine - open source ('GEOS'). rgeos. 2023. URL: https://rgeos.r-forge.r-project.org/ [accessed 2023-01-18]
60. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York, NY. Springer; 2009.
61. Bivand RS, Pebesma E, Gomez-Rubio V. Applied Spatial Data Analysis with R. New York, NY. Springer; 2013.
62. Jacoby WG. Loess: a nonparametric, graphical tool for depicting relationships between variables. Elect Stud. Dec 2000;19(4):577-613. [FREE Full text] [doi: 10.1016/S0261-3794(99)00028-1]
63. Benach J, Vives A, Amable M, Vanroelen C, Tarafa G, Muntaner C. Precarious employment: understanding an emerging social determinant of health. Annu Rev Public Health. 2014;35:229-253. [doi: 10.1146/annurev-publhealth-032013-182500] [Medline: 24641559]
64. Lozano-Fuentes S, Elizondo-Quiroga D, Farfan-Ale JA, Loroño-Pino MA, Garcia-Rejon J, Gomez-Carro S, et al. Use of Google Earth to strengthen public health capacity and facilitate management of vector-borne diseases in resource-poor environments. Bull World Health Organ. Sep 01, 2008;86(9):718-725. [FREE Full text] [doi: 10.2471/blt.07.045880] [Medline: 18797648]
65. Krotov V, Johnson L, Silva L. Tutorial: legality and ethics of web scraping. Commun Assoc Inf Syst. Dec 2020;47:539-563. [FREE Full text] [doi: 10.17705/1CAIS.04724]

66. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. Presented at: AIES '18; February 2-3, 2018, 2018; New Orleans, LA. URL: https://doi.org/10.1145/3278721.3278729 [doi: 10.1145/3278721.3278729]
67. Watt D. On becoming a qualitative researcher: the value of reflexivity. Qual Report. 2007;12(1):82-101. [FREE Full text] [doi: 10.46743/2160-3715/2007.1645]
68. Nilsen P, Bernhardsson S. Context matters in implementation science: a scoping review of determinant frameworks that describe contextual determinants for implementation outcomes. BMC Health Serv Res. Mar 25, 2019;19(1):189. [FREE Full text] [doi: 10.1186/s12913-019-4015-3] [Medline: 30909897]
69. Boegershausen J, Datta H, Borah A, Stephen AT. Fields of gold: scraping web data for marketing insights. J Mark. Aug 02, 2022;86(5):1-20. [doi: 10.1177/00222429221100750]

Abbreviations

BHA: basic health area
CRS: coordinate reference system
SDOH: social determinants of health
WeTMS: web scraping, text mining, and spatial overlay analysis

Edited by A Mavragani; submitted 28.06.23; peer-reviewed by N Levitz, L Rashidian; comments to author 09.10.23; revised version received 20.11.23; accepted 28.11.23; published 08.01.24

Please cite as:
Galvez-Hernandez P, Gonzalez-Viana A, Gonzalez-de Paz L, Shankardass K, Muntaner C
Generating Contextual Variables From Web-Based Data for Health Research: Tutorial on Web Scraping, Text Mining, and Spatial Overlay Analysis
JMIR Public Health Surveill 2024;10:e50379
URL: https://publichealth.jmir.org/2024/1/e50379
doi: 10.2196/50379
PMID: 38190245