

Web scraping considerations

Data Science in a Box
datasciencebox.org



Ethics



datasciencebox.org

"Can you?" vs "Should you?"

Researchers just released profile data on 70,000 OkCupid users without permission

By Brian Resnick | @B_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

A group of researchers has released a data set on nearly 70,000 users of the online dating site OkCupid. The data dump breaks the cardinal rule of social science research ethics: **It took identifiable personal data without permission.**

The information — while publicly available to OkCupid users — was collected by Danish researchers who never contacted OkCupid or its clientele about using it.

The data, collected from November 2014 to March 2015, includes user names, ages, gender, religion, and personality traits, as well as answers to the personal questions the site asks to help match potential mates. The users hail from a few dozen countries around the world.

The data dump did not reveal anyone's real name. But it's entirely possible to use clues from a user's location, demographics, and OkCupid user name to determine their identity.

If your OkC username is one you've used anywhere else, I now know your sexual preferences & kinks, your answers to thousands of questions.

— Scott B. Weingart (@scott_bot) May 11, 2016

Source: Brian Resnick, Researchers just released profile data on 70,000 OkCupid users without permission, Vox.



"Can you?" vs "Should you?"

The image shows a Twitter feed with three tweets. The first tweet is from Emil OW Kirkegaard (@KirkegaardEmil) on May 8, announcing the submission of an OKCupid paper and the availability of its dataset. The second tweet is from Ethan Jewett (@esjewett) on May 11, expressing concern about the dataset's re-identifiability and asking if it was anonymized. The third tweet is a reply from Emil OW Kirkegaard (@KirkegaardEmil) to Ethan Jewett, stating that the data is already public.

Emil OW Kirkegaard @KirkegaardEmil · May 8
The OKCupid paper has now been submitted. This means that the dataset is now public! Enjoy! :) [openpsych.net/forum/showthre...](http://openpsych.net/forum/showthread.php?tid=100)

Ethan Jewett @esjewett · May 11
@KirkegaardEmil This data set is highly re-identifiable. Even includes usernames? Was any work at all done to anonymize it?

**Emil OW Kirkegaard
@KirkegaardEmil**

@esjewett No. Data is already public.



Challenges



Unreliable formatting at the source

Screenshot of a Gumtree search results page for "Used Cars, Vans & Motorbikes" in Edinburgh.

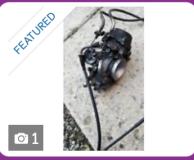
Category:

- < All Categories
- Motors
 - Cars 1,414
 - Parts 563
 - Accessories 505
 - Motorbikes & Scooters 164
 - Vans 142
 - Other Vehicles 47
 - Wanted 24
 - Campervans & Motorhomes 13
 - Caravans 8
 - Plant & Tractors 1

Filters:

Other options

- Urgent ads 3
- Feature ads 37
- Ads with pictures 2,793
- Search title & description

	Suzuki DR650 SE OEM BST40 Carb may fit KTM 640 LC4 Corstorphine, Edinburgh Have upgraded my 1998 DR650SE with the TM40 so surplus is my OEM BST40 comes with throttle cables choke also mic...	£155	29 days ago
	2006 Nissan Micra 1.2 Sport 5dr HATCHBACK Petrol M... Joppa, Edinburgh 2006, NISSAN, MICRA, 1.2 Sport 5dr, 5 Doors, HATCHBACK, Grey, 1895GBP, Petrol, 1240, 60000 V5 Registration Docume...	£1,895	5 mins ago
	Volvo, 960, Saloon, 1996, Automatic, 2922 (cc), 4 doors Liberton, Edinburgh This Volvo 960 is quite simply a lovely car; built when Volvos were real Volvos with a build quality, which in my opinion is second ...	£3,000	1 day ago
	Skoda octavia 2.0TDI 150 DSG Easter Road, Edinburgh Skoda octavia 2016 (66) 2.0 TDI 150 DSG 76 000 miles • Euro 6 • automatic gearbox 6-gear • Sat nav • Half-Leather • Bluetooth...	£7,500	63 days ago

Data broken into many pages

The screenshot shows a web browser displaying a search results page from [yelp.co.uk](https://yelp.co.uk/search?find_desc=Vegetarian&find_loc=Edinburgh&ns=1). The search parameters are set to "Vegetarian" and "Edinburgh". The results are filtered by "Restaurants". The page displays two cards for businesses:

- 9. The Edinburgh Larder**: 4.5 stars, 248 reviews. Category: Delis, Coffee & Tea Shops. Image: A plate of food including what looks like sausages and mushrooms.
- 10. Hanam's**: 4.5 stars, 62 reviews. Category: Middle Eastern. Image: A plate of Middle Eastern cuisine, possibly kebabs or shawarma.

Below the cards is a navigation bar with numbers 1 through 9, followed by a right arrow, indicating there are 24 pages in total. The current page is 1 of 24. To the right of the cards is a map of Edinburgh showing the locations of the top 10 restaurants. Red circles with numbers 1 through 10 indicate the rank of each restaurant from west to east across the city center.

Workflow



datasciencebox.org

Screen scraping vs. APIs

Two different scenarios for web scraping:

- Screen scraping: extract data from source code of website, with html parser (easy) or regular expression matching (less easy)
- Web APIs (application programming interface): website offers a set of structured http requests that return JSON or XML files



A new R workflow

- When working in an R Markdown document, your analysis is re-run each time you knit
- If web scraping in an R Markdown document, you'd be re-scraping the data each time you knit, which is undesirable (and not *nice*)!
- An alternative workflow:
 - Use an R script to save your code
 - Saving interim data scraped using the code in the script as CSV or RDS files
 - Use the saved data in your analysis in your R Markdown document

