

Web scraping considerations

Data Science in a Box

datasciencebox.org



Ethics



"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 screenshot of a Twitter thread. The top tweet is from Emil OW Kirkegaard (@KirkegaardEmil) dated May 8, stating that the OKCupid paper has been submitted and the dataset is now public. The second tweet is a reply from Ethan Jewett (@esjewett) dated May 11, questioning the anonymization of the dataset. The third tweet is a reply from Emil OW Kirkegaard (@KirkegaardEmil) stating that the data is already public. The interface includes profile pictures, names, handles, dates, text, and interaction icons (reply, retweet, like, and a menu).

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...

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

The screenshot shows a Gumtree search results page for 'Used Cars, Vans & Motorbikes' in Edinburgh. The page lists four items with various specifications highlighted by purple boxes to show inconsistent formatting:

- 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...
2006 | 60,000 miles | Trade | Petrol | 1,240 cc (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 ...
1996 | 83,860 miles | Petrol | 2,922 cc (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-Lether • Bluetooth...
2016 | 76,000 miles | Diesel | 1,968 cc (63 days ago)

The left sidebar contains navigation and filter options:

- 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



Data broken into many pages

The screenshot shows a web browser window with the URL `yelp.co.uk/search?find_desc=Vegetarian&find_loc=Edinburgh&ns=1`. The page displays search results for vegetarian restaurants in Edinburgh. Two results are visible:

- 9. The Edinburgh Larder**: 4.5 stars, 248 reviews, ££, Delis, Coffee & Tea Shops.
- 10. Hanam's**: 4.5 stars, 62 reviews, ££, Middle Eastern.

A pagination bar at the bottom of the results list shows numbers 1 through 9, with '1' highlighted in a purple box. To the right of the pagination bar, it says '1 of 24'. On the right side of the page, a map of Edinburgh shows the locations of the top 10 results marked with red numbered circles (1-10).



Workflow



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

