Scraping for stories

@PaulBradshaw
Course leader, MA Data Journalism, Birmingham City University
Data journalist, BBC Shared Data Unit

You are planning a story...

1. ...on zero hours contracts with agencies. How can you establish just **how many agency jobs are advertised as full time**?
2. ...on **how long prisons were closed** and children unable to see their parents while the rest of the country was out of lockdown and able to meet - how?
3. ...**on discrimination in housing that prevents people finding rooms to rent** - but how do you establish how big the problem is?

https://www.youtube.com/watch?v=Efr-VEkwWoM

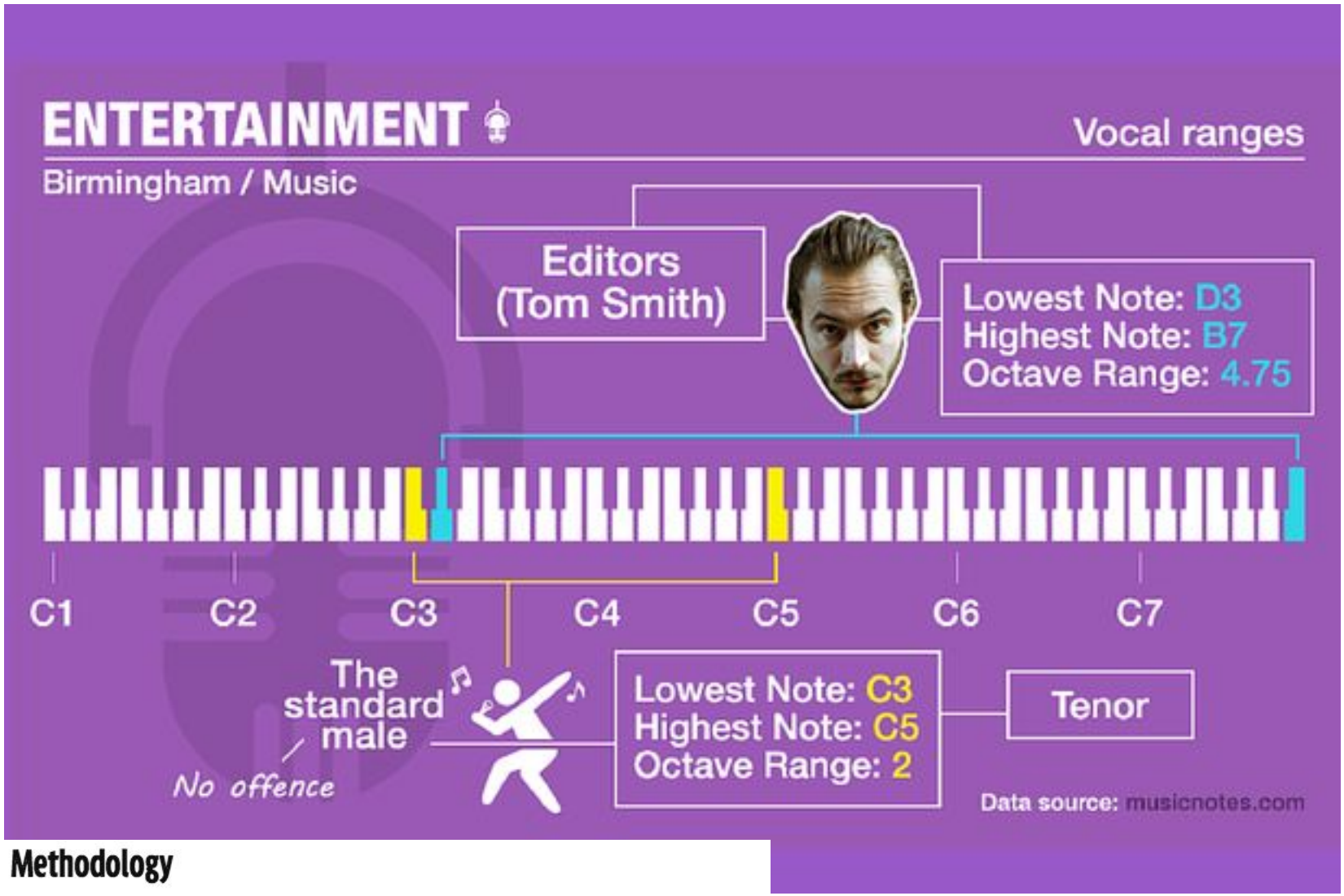
What is scraping?

Automating the gathering of information from **online sources**

Typically, multiple **pages, documents or updates**

...across **space** (e.g. a website) **or time** (e.g. every day or hour).

Which singer has the best vocal range in the UK - No, it's not who you think

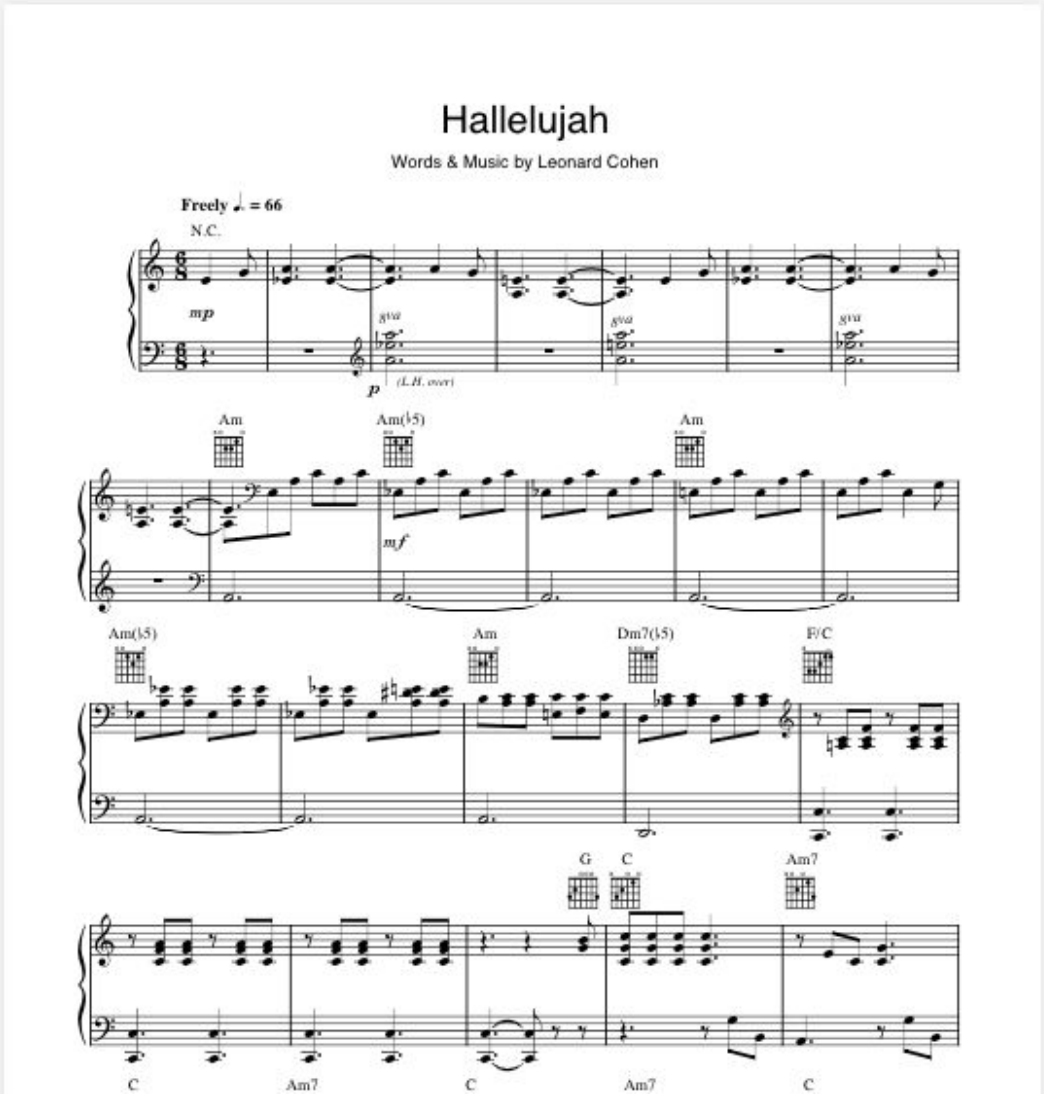


Methodology

In case you're wondering, this list is based on sheet music data from musicnotes.com, a website with over 260,000 sheet music arrangements. We took vocal melodies from UK bands and artists and plotted the highest and lowest notes each singer has recorded.

http://www.mirror.co.uk/news/uk-news/singer-best-vocal-range-uk-4323076

Print Sample | 1 of 6



Hallelujah

BY JEFF BUCKLEY - DIGITAL SHEET MUSIC

Price: £4.20
Save Up To 25% with Volume Discounts

Includes digital copy + 1 print.
Each additional print is £3.04

Transpose (9)

See other arrangements of this song

QUICK DETAILS

Scoring:	Piano/Vocal/Guitar
Instruments:	Voice, range: C4-F5 • Piano • Guitar
Pages:	6
Avg. Rating:	★★★★½
Product #:	MN0053340
Lyrics:	Contains complete lyrics

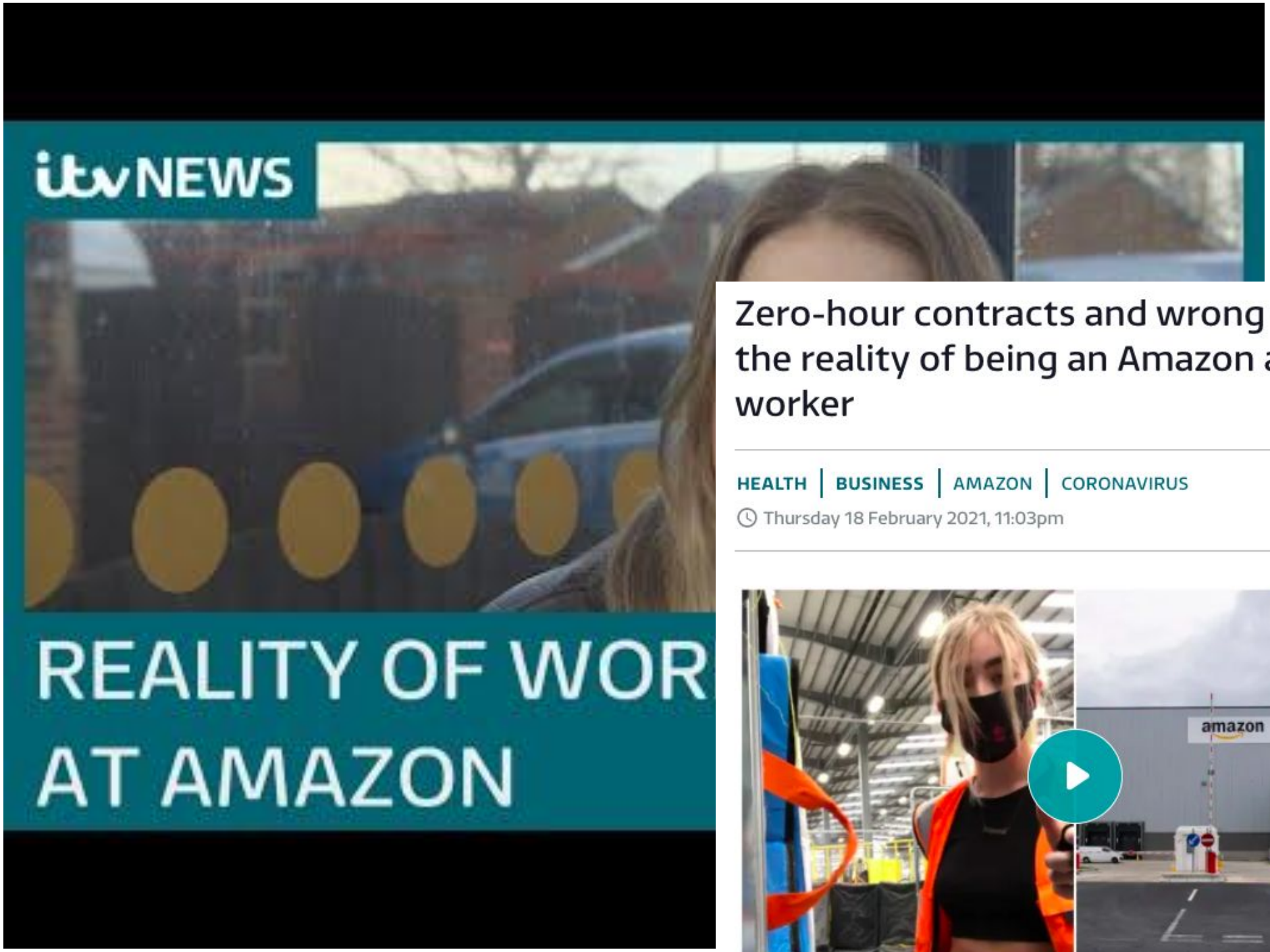
View Full Product Details

Why scraping?

1. Proof
2. Context
3. Exclusivity
4. Leads

Why scraping?

1. **Proof**
2. Context
3. Exclusivity
4. Leads



Zero-hour contracts and wrong wages - the reality of being an Amazon agency worker

HEALTH | BUSINESS | AMAZON | CORONAVIRUS

Thursday 18 February 2021, 11:03pm

https://www.youtube.com/watch?v=xZfxWGZdHMs

Matt Chase for BuzzFeed Ne

THE TENNIS RACKET

No DSS: Most flat shares refuse benefit claimants

9 March 2017 | Comments

Benefit claimants face landlord discrimination despite ruling

By Paul Bradshaw & Alex Homer
BBC Shared Data Unit

28 August 2020

https://github.com/search?q=topic%3Adss+org%3ABBC-Data-Unit&type=Repositories

Leicester House Share - Refurbished & Bills Paid

Ad Details | Message

New Ad ref# 8549557 Mark as unsuitable Save Share

The video gives an accurate reflection of the house as it currently is. NOTE, THE ROOM AVAILABLE IS NOT THE ROOM AT THE TOP OF THE STAIRS, the two rooms available are ground floor rooms. See photos for actual room. If you are interested having viewed the video then we can now arrange a viewing. A deposit will be required to secure the room.

We have NO HIDDEN FEES - A house which has been

House share
Leicester
LE3

£395 pcm (double)
£395 pcm (double)

Availability

Available	Now
Minimum term	6 months
Maximum term	None

Extra cost

Deposit (Room 1)	£350.00
Deposit (Room 2)	£350.00
Bills included?	Yes

Amenities

Furnishings	Furnished
Parking	No
Garage	No

New housemate preferences

Couples OK?	No
Smoking OK?	No
Pets OK?	No
Occupation	Professional
References?	Yes
Min age	18
Gender	Males or females

https://github.com/search?q=topic%3Adss+org%3ABBC-Data-Unit&type=Repositories

Why scraping?

1. Proof
2. **Context**
3. Exclusivity
4. Leads

Coronavirus: Prisoners' children 'forgotten' during pandemic

By Paul Lynch & Paul Bradshaw
BBC Shared Data Unit

30 March

Coronavirus pandemic

Facilities were not given the all-clear to reopen visiting halls again until 6 July - providing the MoJ approved the safety measures put in place at individual sites.

The BBC has found that, while more than half of prisons resumed visits by the end of July, some 5,000 inmates had to wait until September or later.

Among them, HMP Leicester did not reopen until 26 October, just 11 days before it was locked down again.

https://www.bbc.co.uk/news/uk-56420186
https://github.com/BBC-Data-Unit/prisons-children-coronavirus

Heathrow Airport noise complaint every five minutes

By Daniel Wainwright
BBC News

1 November 2016

Heathrow Airport expansion

https://github.com/BBC-Data-Unit/Heathrow-noise
https://onlinejournalismblog.com/2016/11/29/how-the-bbc-england-data-unit-scraped-airport-noise-complaints/

https://www.youtube.com/watch?v=DWRGqmywNYs&t=160s

https://www.youtube.com/watch?v=5A9c6Yyw47k

Why scraping?

1. Proof
2. Context
3. **Exclusivity**
4. Leads

Give MPs deadline on hiring relatives, campaigners urge

By Pete Sherlock and Paul Bradshaw
BBC News

31 July 2017 | Comments

One in five MPs continue to employ a member of their family using taxpayers' money despite the practice being banned for new members of Parliament.

Of the 589 returning MPs, 122 have declared the employment of a relative in the **latest Register of Members' Financial Interests.**

None of the 61 new MPs who secured their seats at the general election on 8 June are allowed to do so.

https://www.bbc.co.uk/news/uk-england-40709220
https://github.com/BBC-Data-Unit/mps-registers-of-interest

Why scraping?

1. Proof
2. Context
3. Exclusivity
4. **Leads**

Olympic Torchbearers

8,000 inspirational people carried the Olympic Flame across the UK. Nominated by someone they knew, it was their moment to shine, inspiring millions of people watching in their community, in the UK and worldwide.

Spotlight

Daley Thompson

Nathan Robertson

Steve Backley

Search Torchbearers

Date | Location

Select date

Search Torchbearers

Date | Location

Select date

The INDEPENDENT

parental control and Digital Vault

NEWS | OPINION | SPORT | OLYMPICS | LIFE | PROPERTY | ARTS & ENTS | TRAVEL | MONEY

Athletics | Cricket | Football | Golf | Motor Racing | Olympics | Racing | Rugby League | Rugby Union | Sailing | Tenn

Hot Topics | Euro 2012 | Wimbledon | Olympics

Olympics Sport > Olympics

Seb Coe promised an 'uplifting torch relay to inspire a generation'. So are these really the role models to do it?

The intention was that 8,000 local heroes with tales to motiv
carry the flame around Britain. But the 2012 sponsors had th

london2012

Mail Online

Home | News | U.S. | Sport | TV&Showbiz | Femail | Health | Science | M

RightMinds Home | News Board | Sport Boards | Showbiz Boards | Femail Boards | He

Flames of Greed: How the Olympic s been hijacked by commercialism

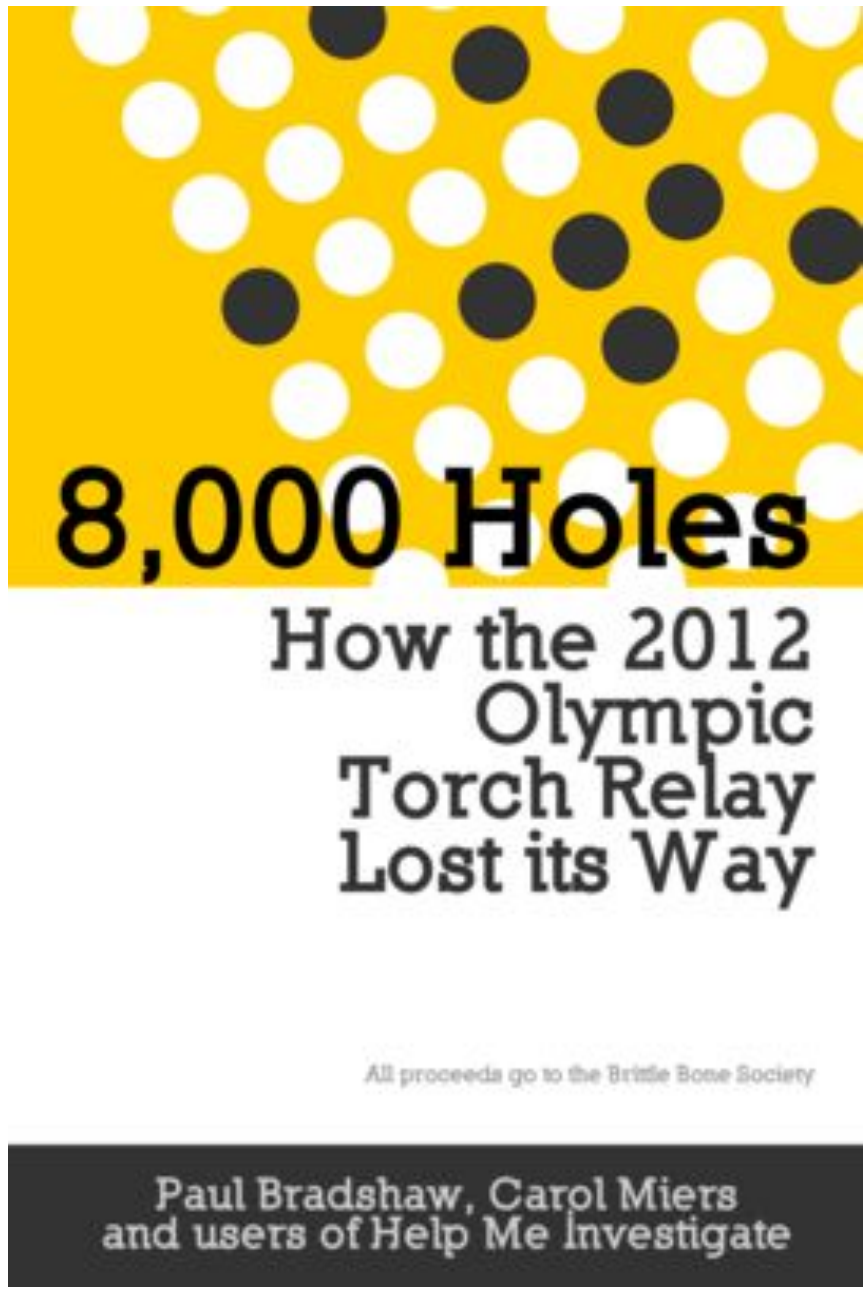
By HARRY MOUNT

London 2012: torchbearers picked by sponsors keep flame of commerce alive

Companies sponsoring the Games have been awarded hundreds of slots in the torch relay. Their nominations include directors, clients – and a steel billionaire

James Ball
guardian.co.uk, Wednesday 6 June 2012 14.09 BST
Comments (28)

leanpub.com/8000holes



Tools

1. OutWit Hub (e.g. job ads, tables on webpages)
2. Octoparse
3. Google Sheets: IMPORTXML function
4. Code: Python or R or JavaScript or command line, etc. (Hacks/Hackers - find a developer)

leanpub.com/u/paulbradshaw
@paulbradshaw
onlinejournalismblog.com
helpmeinvestigate.com
slideshare.net/onlinejournalist
linkedin.com/in/onlinejournalist