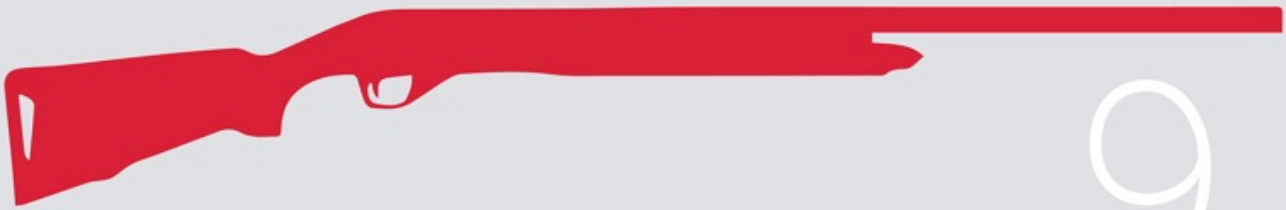
DATA JOURNALISM HEIST

How to get in, get the data, get the story out – and make sure nobody gets hurt!



PAUL BRADSHAW

ONLINE JOURNALISM BLOG

A conversation.

Data Journalism Heist

How to get in, get the data, and get the story out - and make sure nobody gets hurt

Paul Bradshaw

This book is for sale at http://leanpub.com/DataJournalismHeist

This version was published on 2015-06-10

Tweet This Book!

Please help Paul Bradshaw by spreading the word about this book on Twitter!

The suggested hashtag for this book is #djheist.

Find out what other people are saying about the book by clicking on this link to search for this hashtag on Twitter:

https://twitter.com/search?q=#djheist

Also By Paul Bradshaw

Scraping for Journalists

8000 Holes: How the 2012 Olympic Torch Relay Lost its Way

Model for the 21st Century Newsroom - Redux

Stories and Streams

Organising an Online Investigation Team

Finding Stories in Spreadsheets

Excel para periodistas

Learning HTML and CSS by making tweetable quotes

Contents

Huh?

> ***Rusty**: You'd need at least a dozen guys doing a combination of cons.*
>
> ***Danny**: Like what, do you think?*
>
> ***Rusty**: Off the top of my head, I'd say you're looking at a Boeski, a Jim Brown, a Miss Daisy, two Jethros and a Leon Spinks, not to mention the biggest Ella Fitzgerald ever."*
>
> *Ocean's Eleven* (2001)

Can you learn data journalism in an hour?

That was the challenge I was set in late 2011, when I was invited to Bristol to deliver a short workshop. There was a certain appeal in the challenge: there is a myth that data journalism has to be complicated, spectacular, or resource-intensive. But data journalism is not always like that.

For every headline-grabbing Wikileaks story[1] or MPs' expenses saga[2], there are dozens of everyday uses of data journalism that go unnoticed. It might be working out who's the top-scoring Englishman in the Premier League, or seeing whether there's been an outbreak of flu in your area. It might be finding out the worst performing schools, or that season's biggest fashion trends.

So I stripped back everything to some basic techniques. This book covers the bare bones of data journalism: the basic skills to do those simple stories - from finding data in the first place, to getting to the story you want quickly, to following it up and telling it well.

Can you learn data journalism in an hour? Not all of it. But you can learn enough to get started, and get your first stories. More importantly, you can learn enough to see what's possible, with results that provide a basis to begin to learn more (I'll talk about places to go next at the end).

So this is the 'Data Journalism Heist': nothing illegal, but rather a concept designed to reinforce the rough and ready, fast and clean aspect of this approach - as well as the importance of the last part: '*No one gets hurt*'.

It is a book all about speed, and we're wasting time. Time to get started.

[1] http://www.theguardian.com/world/iraq-war-logs

[2] http://www.telegraph.co.uk/news/newstopics/mps-expenses/

The scouting mission: where's the data?

Every good heist begins with a 'recce': a reconnaissance mission to check out the site of our operation. Data journalism has a number of key sites to 'recce':

- If your government, region or city has an open data portal, that should have regular updates. You can find a list at CKAN[3]: open data sites range from the UK's data.gov.uk[4], to regional and city sites like Waterloo[5] in Canada, Emilia-Romagna[6] in Italy, or Chicago[7] in the US;
- If there's an office of national statistics, sign up for updates. One list of national statistical bodies can be found on Wikipedia[8];
- FOI requests are often a good source of data. If you have a site in your country that allows people to send and monitor these (such as Whatdotheyknow.com[9] in the UK and AskTheEU[10] for EU-related FOI requests) then these often provide an alert facility. Also look for specific bodies' disclosure logs - where they publish FOI requests received.

Take some time to check these out. It's a good idea to have data regularly come to you - either by email or, if you use an RSS reader to follow feeds from various sites, that.

If the website has neither email nor RSS updates, try using ChangeDetection.com[11] - this will send you an email when a webpage changes.

If you are reporting on - or particularly interested in - a specific field like crime, health, education, welfare or the environment, try to find bodies (local, national and international) that publish data regularly in that area.

Here are just a few general sources of regular data in the UK alone:

[3] http://ckan.org/instances/
[4] http://data.gov.uk
[5] http://www.regionofwaterloo.ca/en/regionalgovernment/OpenDataHome.asp
[6] http://datablog.ahref.eu/en/ahref-log/opendata/open-data-la-regione-emilia-romagna-presenta-il-suo-portale
[7] https://data.cityofchicago.org/
[8] http://en.wikipedia.org/wiki/List_of_national_and_international_statistical_services
[9] http://whatdotheyknow.com
[10] http://www.asktheeu.org
[11] http://ChangeDetection.com

- Education: the Higher Education Statistics Agency (HESA), the Higher Education Funding Council (HEFCE) and Universities and Colleges Admission Service (UCAS) all hold data on universities and students. Ofsted, the regulator of schools, hold data on pupils and education at pre-16 level.
- The NHS Information Centre (NHSIC) and Health Episode Statistics (HES) hold data on hospital admissions and local doctors.
- NOMIS holds data on the labour market: where people are employed and unemployed.
- Data.police.uk holds data on crime and policing

Scouting a local government website

This scouting mission concerns a typical type of data which you might find landing in your alerts regularly: local government spending.

In England every local council is required to publish data on its spending over £500. To find it, search for `expenditure 500` and the name of a local authority. In our case, we're going to look at Birmingham - the biggest local authority in Europe.

> *What the data contains doesn't really matter here: the point is the exercises you'll be going through in looking at it. You can apply the same process to most regular public datasets, whether it's employment data, environmental information, or weather.*

Birmingham City Council's expenditure data is at Birmingham.gov.uk/payment-data[12]. You'll find monthly spending data going back over the last year, and can request older data by email.

It comes in two formats: PDFs, and spreadsheets. Given the choice, always avoid PDFs.

> If you prefer to work on some international data, download loans data from the European Investment Bank[a] - change the search to the widest possible criteria and then use the (easy to miss) *Export* link at the bottom of the page to get an Excel spreadsheet.
>
> [a] http://www.eib.org/projects/loans/list/index.htm

The spreadsheets are shown here as a Microsoft Excel icon, but they are actually **CSV** files which will work on any spreadsheet software, including free options like Google Drive spreadsheets and Open Office[13] (which can also open Excel spreadsheets).

[12] http://www.birmingham.gov.uk/payment-data

[13] http://www.openoffice.org/

CSV stands for Comma Separated Values - this means that the value in each column is separated by a comma like so: `Name, Date of birth, Address`. When the spreadsheet software opens this, it replaces each comma with a new column.

Click on the most recent *spreadsheet* version of the spending data (not the PDF) and download it to your computer. Make sure you save it somewhere you can find later, like your desktop.

Once downloaded, open it in your preferred spreadsheet software - either by double-clicking it or uploading to a web-based tool like Google Drive[14].

Now, we're in.

If PDF is the only option, try a PDF-to-Excel converter like PDFtoExcelOnline.com[15], PDF2XL[16] or Wondershare PDF Converter[17]. You can also try a quick phonecall or FOI request to get the data in spreadsheet format.

Other ways of getting information:

Building contacts is no less important in data journalism than other forms of journalism: these are the people who can alert you to the existence of data, or even leak it to you.

You can also request data using a 'right to information' law, such as the Freedom Of Information Act (FOI). Heather Brooke's book *Your Right To Know* is an essential reference book for that. Finally, you can 'scrape' data using software or programming. See my book *Scraping for Journalists*[18] if you want to learn about that.

[14] https://support.google.com/drive/answer/2424368?hl=en&ref_topic=2375187
[15] PDFtoExcelOnline.com
[16] http://www.cogniview.com/pdf-to-excel/pdf2xl-basic
[17] http://www.wondershare.com/pdf-converter/
[18] http://leanpub.com/scrapingforjournalists