USING DATA SAFELY
Workshop summary | April 28

What stories had we seen?
A bus story in the FT. Might be this one?

What did we cover?

- Data protection
- Other legal areas: licences and copyright
- Securing data
- Data protection as a beat
- Ethical data journalism

Data protection

Has anyone encountered data protection questions at work?

- Can we store all the text messages we receive at a station / to a show in order to analyse this part of the audience?
- Can we use a person's medical data to stand up part of a story?

Where are we with GDPR now that we're not in the EU?

2016-18	(EU) GDPR
2018	Data Protection Act (DPA)
2021	UK GDPR comes into force

As the ICO says "The GDPR is retained in domestic law as the UK GDPR, but the UK has the independence to keep the framework under review." That is, not much has changed for the UK in terms of GDPR, but it might in the future.

What is Special Category data?

As the ICO says, it's personal data tied to

- racial or ethnic origin;
- political opinions;
- religious or philosophical beliefs;
- trade union membership;
- genetic data;
- biometric data;
- health;
- a person's sex life; and
- a person's sexual orientation.

To collect Special Category data you need a good reason: explicit consent from the person is one way, substantial public interest is another. (There are eight other possible ‘good reasons’ laid out by the ICO).

Why do journalists care about data protection?

- Media groups (& freelancers) keep information on people
- Journalists can be prosecuted for how they got (personal) data
- People will refuse information to the media, claiming ‘data protection’
- Data protection is now a patch
- Methodology – we aim to publish our data sources and our methods

So what is personal data?

Essentially, it’s information about an identifiable living individual and that is stored. The ICO outlines it in detail.

Note two things:

1. Personal data pertains to *living* people

2. The *person* concerned is always the *owner* of the data

Subject Access Request

So, because I own my own data, I can ask for the data to be corrected, erased or given to me. This is the basis of what’s called a Subject Access Request.

> If a [person] files a “subject access request” – an email, fax or letter asking for their personal data – the controller will have [about a month] to collate a cache of all the information stored about that person. This includes any email that refers to the worker, as well as performance reviews, job interviews, payroll records, absence records, disciplinary records, computer access logs, CCTV footage, and recordings of phone calls to, from or about the person. (Guardian, 2018)

Journalistic exemptions

Journalists always have to keep it safe but we don’t necessarily have to respect the following data protection requirements:

- lawful, fair and transparent processing
- collection for specified, explicit and legitimate purposes
- material to be adequate, relevant and limited to what’s necessary
- material to be accurate and up to date
- material to permit identification of subject for no longer than necessary

As McNae says:

> Journalists and news organisations holding information for journalistic purposes are not required to respond to questions about whether they hold data or what data they hold, and are not obliged to comply with requests to remove it. The best response to such questions is simply to refuse to respond to them.

The Information Commissioner's Office

The ICO is known to us also as the place to appeal FOI rejections.

It publishes a guide to data protection and the media which is slightly old, but still basically sound.

We can look up any data controller (any person or organisation who has registered with the ICO because they collect and store personal data). The full list of data controllers is available here as a downloadable CSV too, updated daily, and runs to more than a million rows.

There are several BBC bodies listed. The details for the British Broadcasting Corporation give the name of the current DPO (Data Protection Officer) at the BBC. This is ultimately the person at the BBC who understands data protection issues.

We look at the "number of reports of personal data breaches received by the ICO during Q3 2021/22". The CSV file lets us see what categories stand out. These are *reports* however rather than ICO action.

The ICO telephone helpline might be a useful resource if we have questions, although it is for the general public, rather than the press: https://ico.org.uk/global/contact-us/

Obstruction

Obstruction due to perceived data protection *can* be a problem

> "Over the past few years, data protection claims against the media have increased. It appears that they are being deployed as a new form of reputation management: there is no time limitation on claims, no defence for truth or honest opinion, and no requirement to demonstrate serious harm, as there is with libel. Every stage in the practice of journalism is potentially vulnerable to challenge under data protection laws – from researching a story to maintaining archives. Data protection law is, as Keller sees it, 'a powerful new tool for abusive claimants to hide information from the public'." — Glanville

Examples of (possible) obstruction

Dog names are not personal data (but they could be connected to it, via jigsawing?)

Information on benefit claimants and suicide was destroyed, with data protection used as the defence.

Asked on air about the vaccination rate of professional footballers, a PFA representative hid behind ‘data protection’. There is nothing preventing the PFA from saying “x % of professional footballers have been vaccinated”.

It has cropped up with information about public, unelected, workers

> “There have, however, been notable altercations between journalists and councils — particularly in relation to the salaries and perks of senior officers. Councils have often tried to hide behind data protection legislation when asked for such details under FoI, arguing that, because officers are not elected representatives, these details constitute personal information and should be treated as confidential.
>
> The Commissioner has, however, clarified the legal position surrounding this, by distinguishing between information relating to public officials’ private lives (exempt) and public duties (covered). Indeed, a landmark ruling by the Commissioner in 2011 raised the prospect of a more liberal approach to disclosing information about public servants’ salaries. Defying a Cabinet Office attempt to protect the identities of twenty-four senior civil servants earning more than £150,000 each, he ordered that their names be published.”
>
> — Morrison (*Public Affairs*)

Question: Could a photograph of someone, taken legally in a public place, be considered personal data?

Separately (It’s not directly related but in terms of where photos and personal data are headed, it may be relevant): Note the power of the https://pimeyes.com/en website. If you give it a fresh photo of yourself, it can use that to identify photographs of you that exist on the internet. This is not the image search that we’re familiar with, but a machine that collects enough information from a (previously unseen) single photo of you, to find pictures of you online.

https://ipo.blog.gov.uk/2019/06/11/copyright-and-gdpr-for-photographers/
https://ico.org.uk/for-organisations/gdpr-resources/lawful-basis-interactive-guidance-tool/lawful-basis-assessment-report/

OTHER LEGAL POINTS

Copyright

Datasets can be subject to copyright. See the summary of a case around the football fixture lists, in 2012 or OpenStreetMap worried about their databases not getting enough protection after Brexit. A dataset like Baby Names in England & Wales (ONS) carries the terms of its licence on its third sheet in the Excel workbook. Data from the Care Quality Commission has the terms of use laid out on its website.

What does the Guardian and Amazon say about ‘database’ on their pages of Terms and Conditions? Why might we want to gather data from Amazon? (Answer: for stories about price gouging on PPE during the pandemic for example).

Regulatory obligations

We have the same obligations on accuracy when using and communicating numbers. See IPSO ruling (section 11 has the key part) on a Telegraph piece that overplayed a causal relationship in its headline

SECURING DATA

Machines

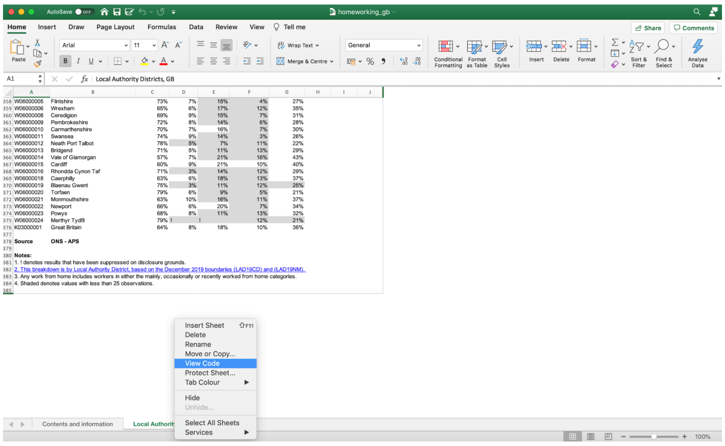
Thought experiment: what data would be lost / exposed if my computer (or indeed, phone) was lost or stolen?

Passwords

You can experiment with the strength of a password using a website, to see how well it holds up against a brute force attack, i.e. the use of a computer to try all the possibilities to crack a password.

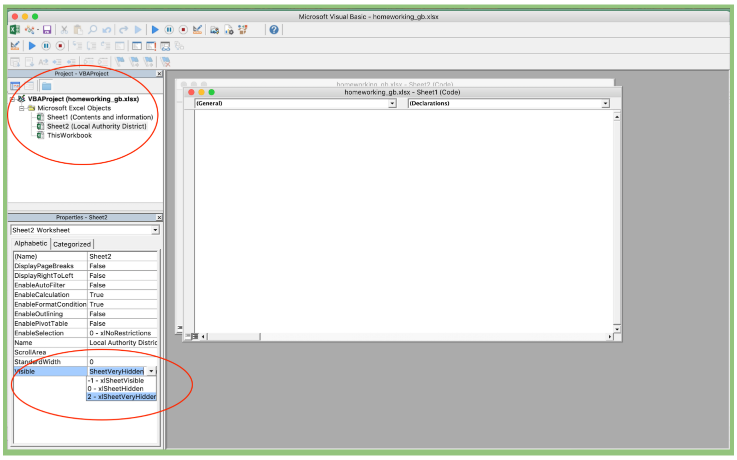
Spreadsheets

- Protect an Excel file with a password from 1. being opened 2. being edited (Usually found under File)
- Protect one Sheet in an Excel workbook with a password (Right click on sheet’s tab)
- Hiding a Sheet has no protective value unless you set it as “Very Hidden”: Right click on the sheet tab / View Code / Select sheet on top left / Select Visible and Very Hidden on bottom left



A1 Local Authority Districts, GB

	A	B	C	D	E	F	G
358	W06000005	Flintshire	73%	7%	15%	4%	27%
359	W06000006	Wrexham	65%	6%	17%	12%	35%
360	W06000008	Ceredigion	69%	9%	15%	7%	31%
361	W06000009	Pembrokeshire	72%	8%	14%	6%	28%
362	W06000010	Carmarthenshire	70%	7%	16%	7%	30%
363	W06000011	Swansea	74%	9%	14%	3%	26%
364	W06000012	Neath Port Talbot	78%	5%	7%	11%	22%
365	W06000013	Bridgend	71%	5%	11%	13%	29%
366	W06000014	Vale of Glamorgan	57%	7%	21%	16%	43%
367	W06000015	Cardiff	60%	9%	21%	10%	40%
368	W06000016	Rhondda Cynon Taf	71%	3%	14%	12%	29%
369	W06000018	Caerphilly	63%	6%	18%	13%	37%
370	W06000019	Blaenau Gwent	75%	3%	11%	12%	25%
371	W06000020	Torfaen	79%	6%	9%	5%	21%
372	W06000021	Monmouthshire	63%	10%	16%	11%	37%
373	W06000022	Newport	66%	6%	20%	7%	34%
374	W06000023	Powys	68%	8%	11%	13%	32%
375	W06000024	Merthyr Tydfil	79%	!	!	12%	21%
376	K03000001	Great Britain	64%	8%	18%	10%	36%
377							
378	**Source**	**ONS - APS**					
379							
380	**Notes:**						
381	1. ! denotes results that have been suppressed on disclosure grounds.						
382	2. This breakdown is by Local Authority District, based on the December 2019 boundaries (LAD19CD) and (LAD19NM).						
383	3. Any work from home includes workers in either the mainly, occasionally or recently worked from home categories.						
384	4. Shaded denotes values with less than 25 observations.						


Microsoft Visual Basic - homeworking_gb.xlsx
Project - VBAProject
VBAProject (homeworking_gb.xlsx)
Microsoft Excel Objects
Sheet1 (Contents and information)
Sheet2 (Local Authority District)
ThisWorkbook
homeworking_gb.xlsx - Sheet2 (Code)
homeworking_gb.xlsx - Sheet1 (Code)
(General)
(Declarations)
Properties - Sheet2
Sheet2 Worksheet
Alphabetic
Categorized

(Name)	Sheet2
DisplayPageBreaks	False
DisplayRightToLeft	False
EnableAutoFilter	False
EnableCalculation	True
EnableFormatCondition	True
EnableOutlining	False
EnablePivotTable	False
EnableSelection	0 - xlNoRestrictions
Name	Local Authority Distric
ScrollArea	
StandardWidth	0
Visible	SheetVeryHidden

-1 - xlSheetVisible
0 - xlSheetHidden
2 - xlSheetVeryHidder

Google Sheets

This workbook has one sheet completely protected (SRC), has *part* of another protected ("Notes" A3:B3), and has another sheet ("Summary") hidden (see View / Hidden Sheets to unhide).

DATA PROTECTION AS A PATCH

Journalism

Data protection can (a little like environment) appear in all sorts of areas: local government, technology, business, sport etc. Review the stories on stories data protection.pdf: What was the source for these stories?

Use of AI (an aside)

The FT story is interesting in the context of data protection because it considers if / how the UK is going to give different levels of protection to citizens (compared to Europeans) particularly when it comes to hidden computational processes.

But the question of artificial intelligence and job applications is interesting in itself. See how a 'black box' in software for evaluating video job applications was investigated by German journalists: https://interaktiv.br.de/ki-bewerbung/en/

Action taken against particular groups for data breaches
See the ICO enforcement list: what did the following groups get wrong?:

- Royal Mail
- Ministry of Justice
- Mermaids

ETHICS

Key point on using data analysis
Taking a dataset, and looking at the data, what does the data say? Nothing.
Data doesn't speak. Data is ***interpreted***.

Ok, fine, but what does that mean practically for us, as journalists?

- When we hear an expert say "The data shows…", they really mean "Our reading of the data shows…"
- An analysis of data might be correct, and the data might be good, but there might be another analysis of it still to be done
- If we hear a politician (or indeed a manager) point to evidence-based decisions by saying 'the data shows that…' we should remember that they mean 'our best reading of the data shows that…' or 'when *we* look at the data, we see that…'
- When we analyse data we can be distracted by being personally invested in the phenomenon being measured. See the first chapter in Tim Harford's *How to make the world add up* on having our analysis misled by our emotions.

Data quality

If the data is no good, the finding won't be either.

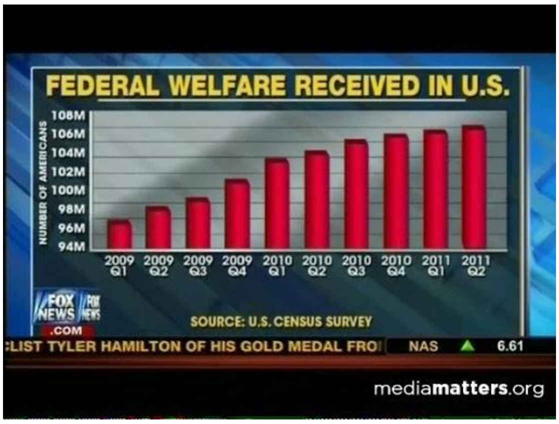
See the 538 story on kidnappings in Nigeria which (now) outlines what went wrong. Note also that nobody was phoned about the 'findings' before publication.

Visualisation

Visualisation is an ethical, not a legal question. When visualisation is misleading, remember that the data is *correct*. It's the *presentation* of it that is erroneous.

Consider, for example

- the mapping migrants question (see How not to map migrant movement from previous module)
- if colour and gender is a problem Pink v Blue (FT)
- how a trend is exaggerated with a cropped vertical axis



As raised in the workshop, are logarithmic scales a good idea? These researchers at LSE suggest they aren't, but lots of good journalism used them during the pandemic. You can try them out on datawrapper, if you have data that works:

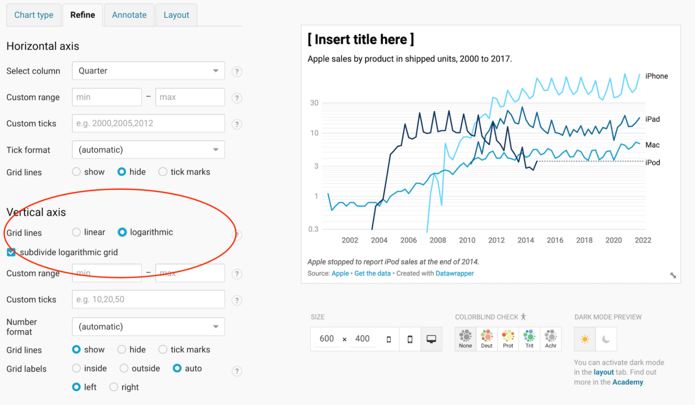


Chart type | Refine | Annotate | Layout

Horizontal axis

Select column: Quarter

Custom range: min – max

Custom ticks: e.g. 2000,2005,2012

Tick format: (automatic)

Grid lines: ○ show ● hide ○ tick marks

Vertical axis

Grid lines: ○ linear ● logarithmic

☑ subdivide logarithmic grid

Custom range: min – max

Custom ticks: e.g. 10,20,50

Number format: (automatic)

Grid lines: ● show ○ hide ○ tick marks

Grid labels: ○ inside ○ outside ● auto

● left ○ right

[Insert title here]

Apple sales by product in shipped units, 2000 to 2017.

Apple stopped to report iPod sales at the end of 2014.

Source: Apple • Get the data • Created with Datawrapper

SIZE: 600 × 400

COLORBLIND CHECK: None, Deut, Prot, Trit, Achr

DARK MODE PREVIEW

You can activate dark mode in the **layout** tab. Find out more in the **Academy**.

Anonymisation

Publishing our analysis or data we've collected is to be encouraged. Sometimes the way to do that is to anonymise the data. But this doesn't always work. Here are two cases where personal data wasn't protected properly.

- Buzzfeed tennis story
- Sandy Hook mapping

Numbers

Some basic tools for ensuring our output is safe when using calculations

- Get another pair of eyes (this could range from a colleague to an in-house subject expert or an external expert)
- Double check our sums afterwards with calculators (lots of percentage calculators online)
- Do rough estimates mentally to see if our figures ring true