Streaming web data for social science research

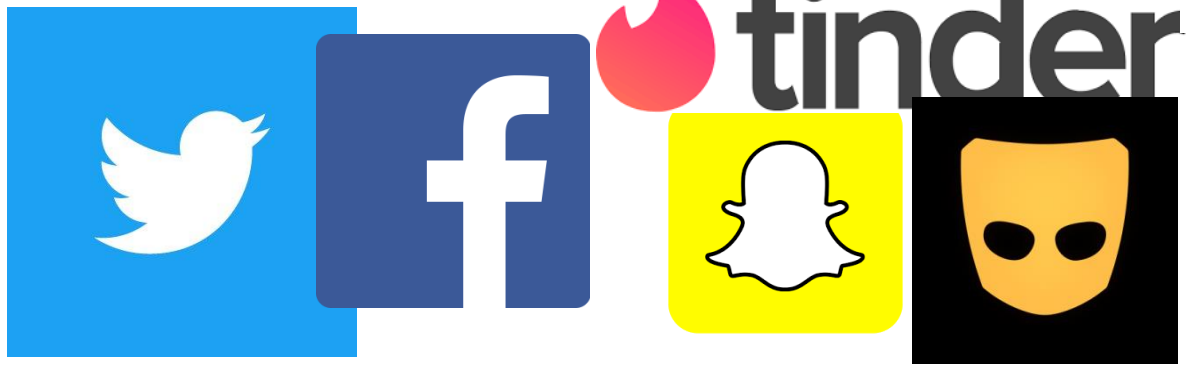
Jonathan Bright
Senior Research Fellow, Oxford Internet Institute
University of Oxford
@jonmbright | jonathan.bright@oii.ox.ac.uk

Overview

- What is (streaming) web data & how is it being used in social science research?
- What are the implications / challenges for research data policy?

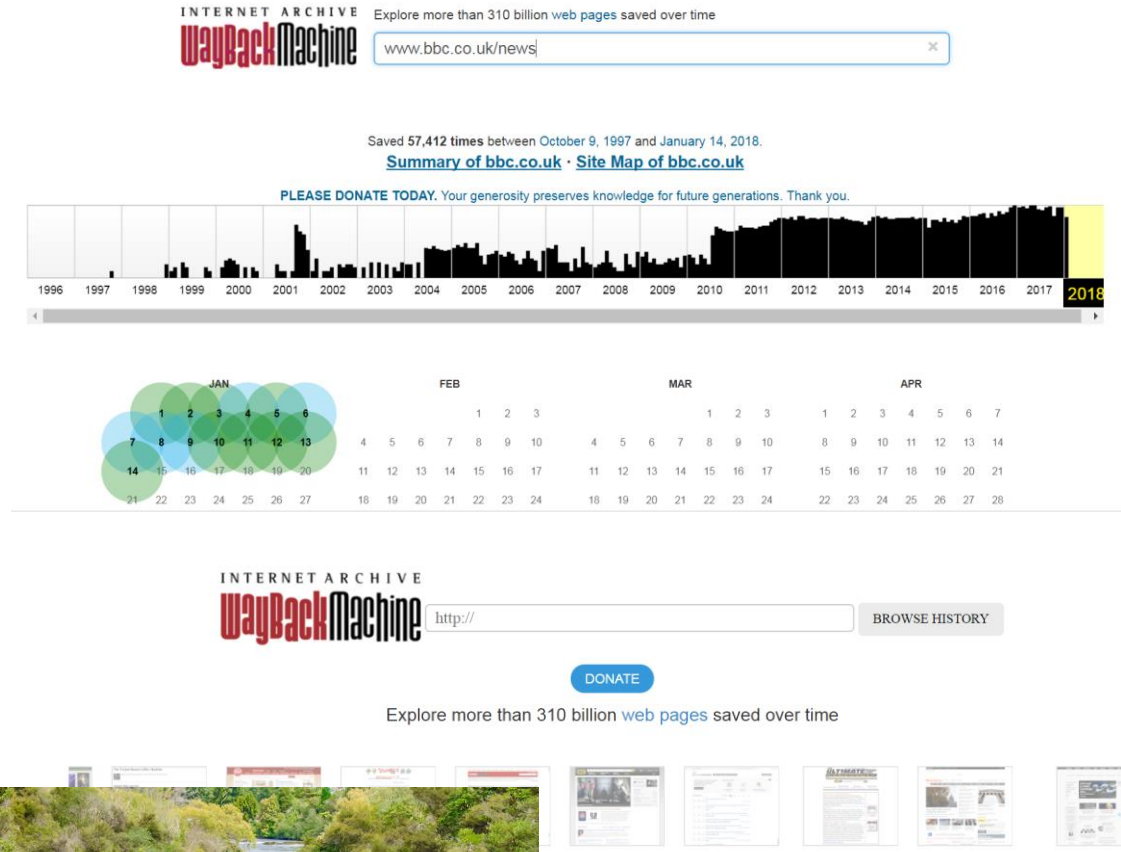
Streaming Web Data and Social Sciences

- Huge amount of social life now takes place online / on an app:
  - Friendship and romantic relationships, job seeking and working, travel and vacations, information consumption, entertainment, political participation, etc.
  - Part of the study of social life is therefore necessarily about the observation of online behaviour!
  - And the web also facilitates this observation



Streaming Web Data and Social Sciences

- One characteristic of the web is that it is constantly changing
  - Most websites will change the content of their front page from one day to the next
  - Social platforms will change when users interact
  - Often things are either “demoted” or “deleted”
- Data on a platform is a bit like water in a river – it “streams” past and is characterised by transience



Streaming Web Data and Social Sciences

- Streaming web data in general has enabled an enormous amount of social science over the past decade
  - E.g. mapping of information diffusion patterns on Twitter, social influence on political participation on Facebook, rare events such as tipping points and power users
- A couple of quick examples from my own work…

Bright, J. 2016. The Social News Gap: How News Reading and News Sharing Diverge. *Journal of Communication*, 66, 343-365.

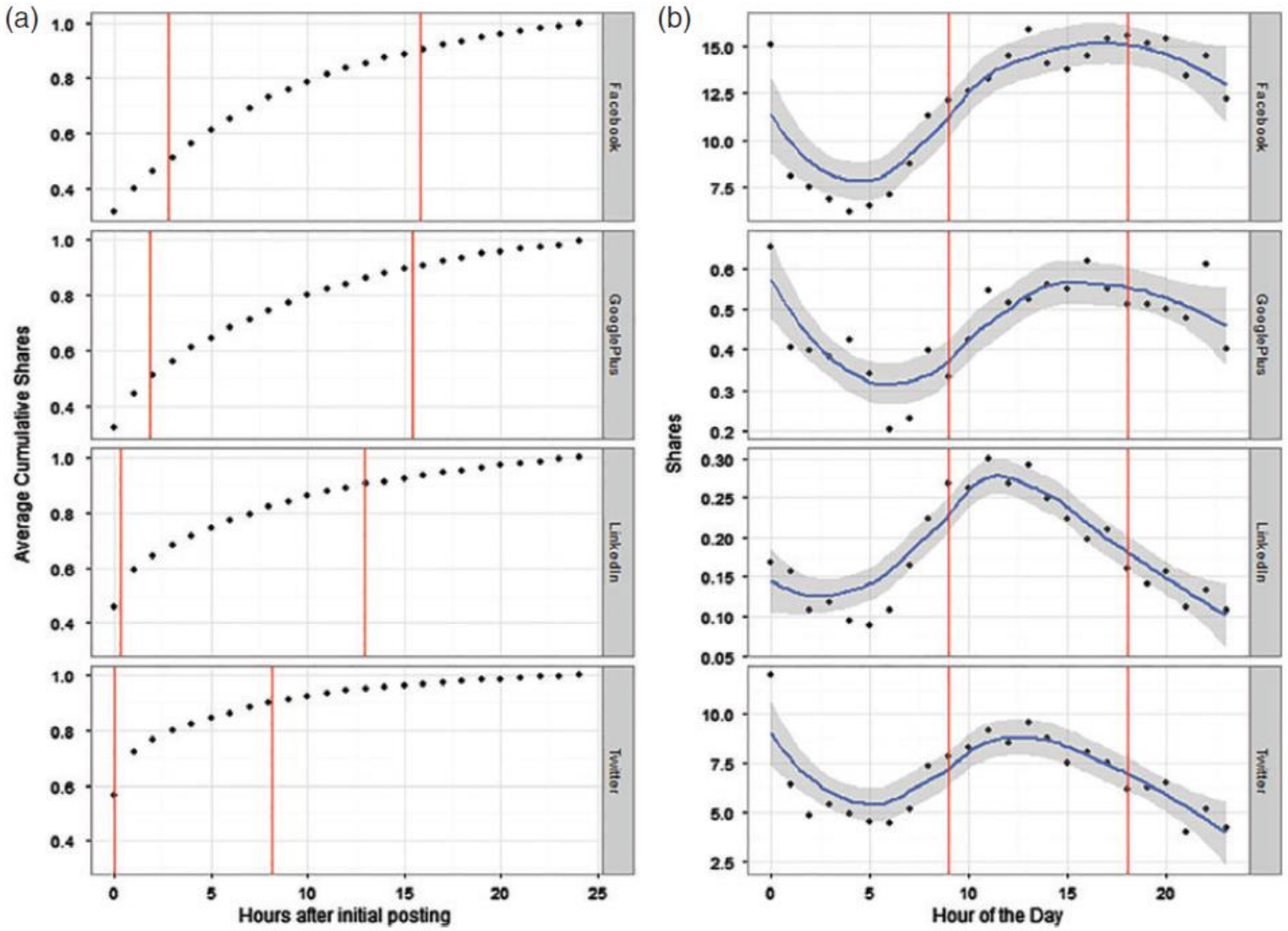
Evolution of sharing

- It is easy to find out the number of times a given online news article (or any piece of online content) has been shared -> simply give the link to the API
- But it's also interesting to observe how sharing evolves after publication, and how this relates to other dynamic factors such as the positioning of the article
- This requires streaming observations:
  - Capturing the front page of the news website at regular intervals to find new articles and determine positional fluctuations of existing articles
  - Querying the API at regular intervals with a dynamically updated list of article links
  - i.e. streaming states of multiple different social systems

Bright, J. 2016. The Social News Gap: How News Reading and News Sharing Diverge. *Journal of Communication*, 66, 343-365.

Results

- Sharing happens very quickly after publication
  - Different patterns on different sites
  - More or less viral?
- Sharing fluctuates predictably throughout the day
  - Again different patterns
  - Different usages?
- Position makes a diff but more important for FB / LinkedIn



Audience Metrics

Bright, J & Nicholls, T. 2014. The Life and Death of Political News: Measuring the Impact of the Audience Agenda Using Online Data. *Social Science Computer Review*, 32(2), 170-181.

- Audience metrics could influence editorial processes, as editors may decide to leave articles on the front page if they are more popular
- Research has looked at whether this is the case by asking if articles which lasted the longest on the front page ever appeared on the “most read” list of articles
  - But -> obvious direction of causality prob
- Requires a streaming approach which tracks whether being on the most read list at time *t* improves the chance of still being on the front page at time *t + 1*
  - Need to capture front page of news website at regular intervals

Results

- Good evidence for metrics effect
- Most read articles last 3 hours longer on the front page
- They are about 25% less likely to be deleted
- Same effect for political and entertainment news
- Effect not present for the Daily Mail

Bright, J & Nicholls, T. 2014. The Life and Death of Political News: Measuring the Impact of the Audience Agenda Using Online Data. *Social Science Computer Review*, 32(2), 170-181.

Paper	Articles	Most Read Slots	Most Read	% Most Read	Median Time to Most Read, hr	Mean Duration (most read articles)	Mean Duration (all other articles)
BBC	6,745	10	1,222	18%	0.25	18	11
Daily Mail	14,484	Up to 20	1,593	11%	2.3	17	20
Guardian	9,380	5	805	9%	2.8	18	13
Mirror	5,075	10	593	12%	4	14	9
Telegraph	7,296	10	711	10%	2.5	20	14
Total	42,980		4,924	12%	2	18	15

	Model 1—Overall Impact of Most Read	Model 2—Political News	Model 3—Entertainment News	Model 4—Quality Papers	Model 5—Tabloid Papers
Most Read at time *T*	0.74 (0.02) ***	0.70 (0.03) ***	0.68 (0.05) ***	0.54 (0.03) ***	1.10 (0.03)**
Entertainment	0.89 (0.01) ***			0.91 (0.01) ***	0.79 (0.01) ***
Daily Mail	0.49 (0.01) ***	0.58 (0.02) ***	0.41 (0.03) ***		
Guardian	0.64 (0.02) ***	0.61 (0.02) ***	0.66 (0.03) ***		
Mirror	1.32 (0.02) ***	1.06 (0.02)*	1.57 (0.03) ***		
Telegraph	0.70 (0.02) ***	0.68 (0.02) ***	0.72 (0.03) ***		
N	2,429,940	1,247,210	1,182,730	1,147,878	1,282,062

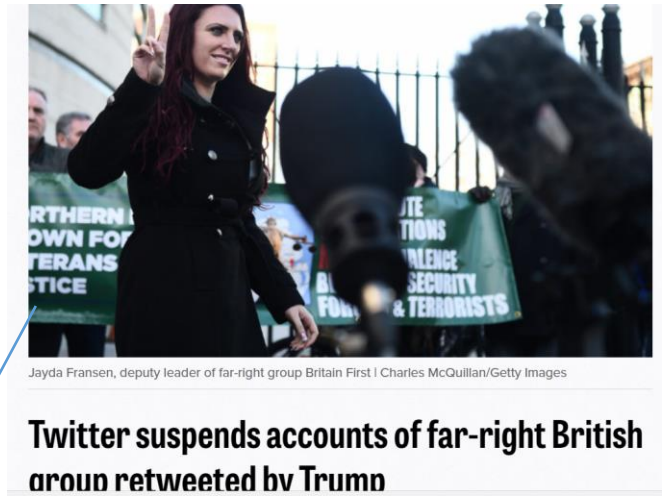
So – what are the challenges for research data and policy?

Challenges

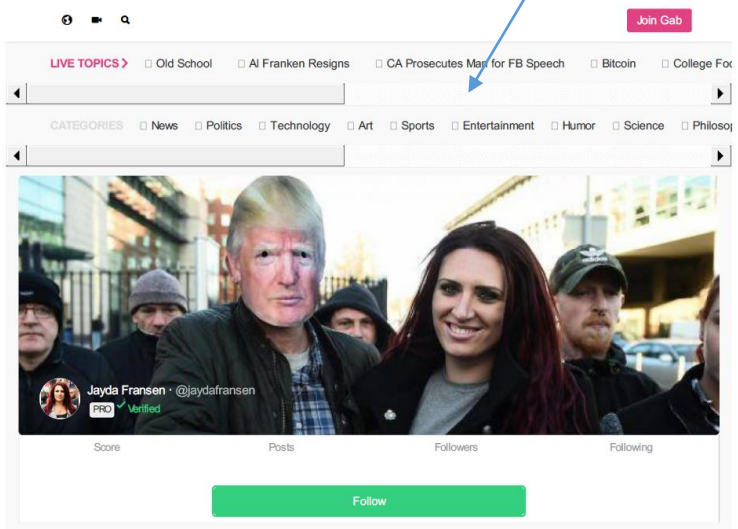
- Previously social science have been characterised by data scarcity
  - Small amounts of large scale research efforts (census, recurring population surveys, etc.)
- This scarcity has had a huge structuring effect on the field
  - Large scale research grants with long time horizons
  - Considerable build up and decommissioning phases (with data preservation built in)
  - Lots of our statistics are mostly oriented towards maximising statistical power of small samples! (at the expense of lots of assumptions)
- But now we are moving into a period of “data abundance”
  - How does this change the paradigm?
  - Some examples…

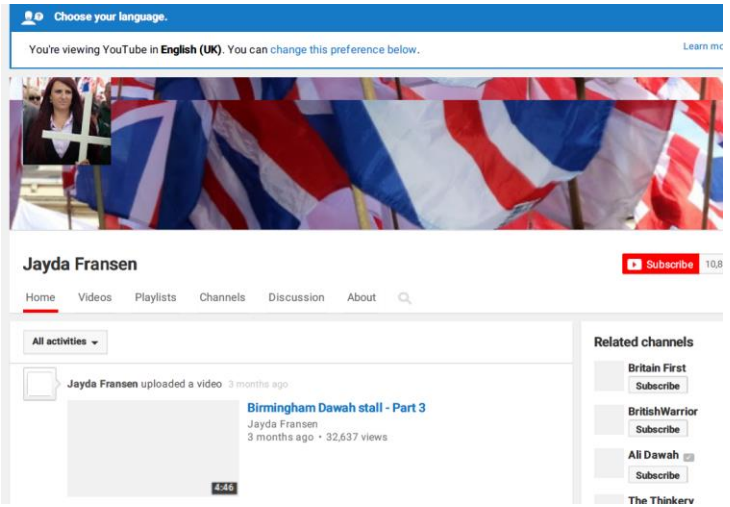
Streaming research cycle

- Streaming data requires you to move really quickly!
- Once things are passed they are lost
- So you need existing data capture tools before events
- Conflicts with grants, ethics committee, IT support, etc.



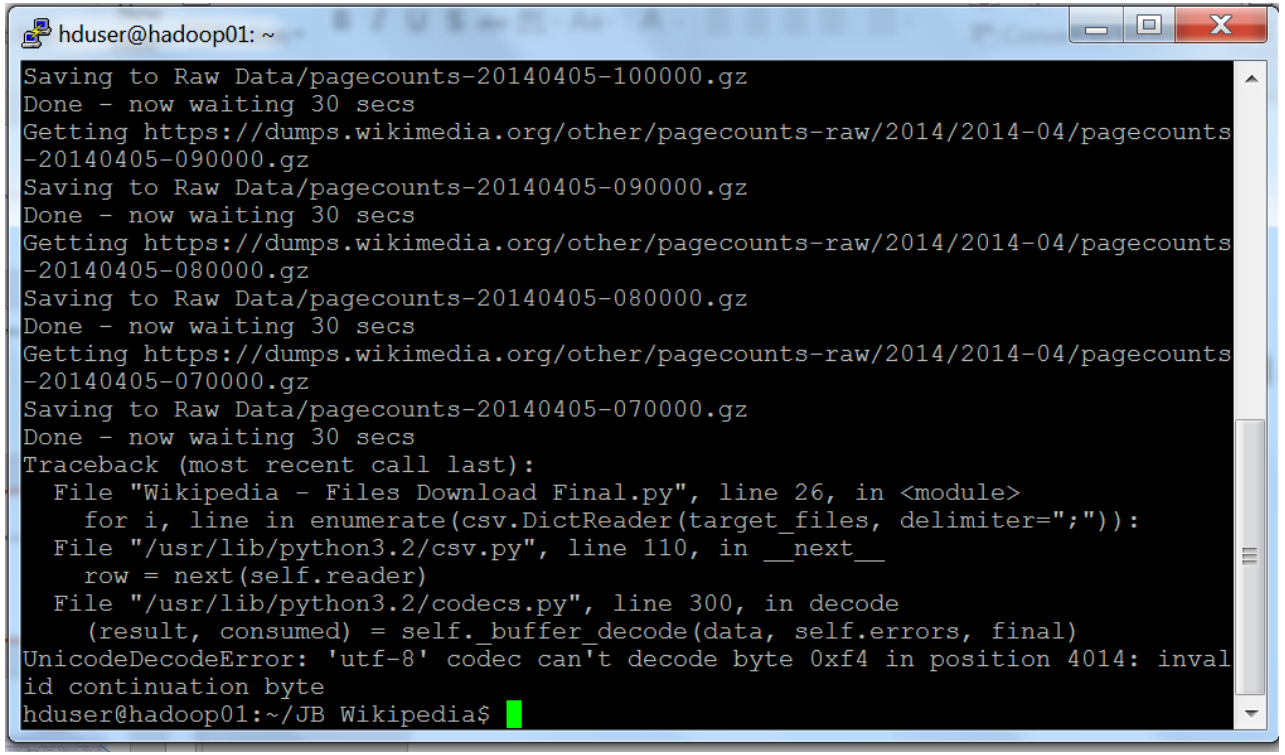

```
{"created_at": "Tue Jan 02 17:40:27 +0000 2018", "id": 948247630987055104, "id_str":
"948247630987055104", "text": "RT @CMilrs: Paul Golding and Jayda Fransen of
#BritainFirst on
Gab.\nhttps://t.co/CRO8TUjUEt\nhttps://t.co/AbRUresOMc\nhttps://t.co/1vNiRmQKvl\u2026",
"truncated": false, "entities": {"hashtags": [{"text": "BritainFirst", "indices": [46,
59]}], "symbols": [], "user_mentions": [{"screen_name": "CMilrs", "name": "SMiles", "id":
```



Facilities

- I need “always on” computing which I can manage from my laptop
- I need that to be secure but also shareable with other researchers on the team
- I need to record date and time of capture etc. and make sure things happen at the right time
- Would it be better to have cloud computing or work with internal IT?



Space

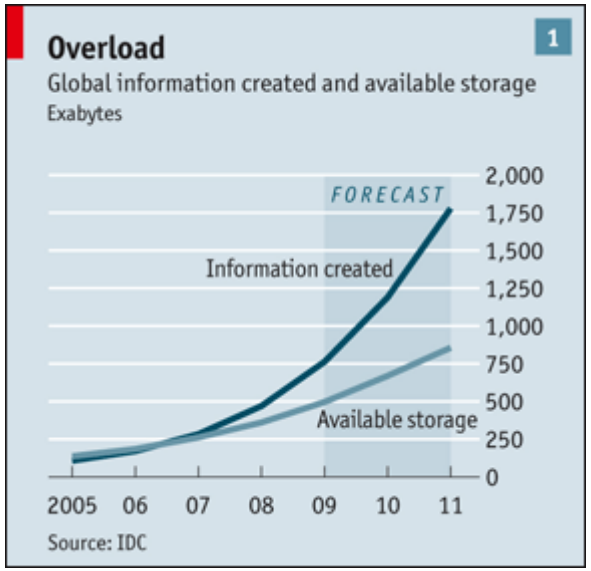
- Streaming data can hog huge amounts of space
  - Can easily generate several GB of tweets per day
- Can we put these on university file systems?
  - Serviced and backed up data warehouses can have a considerable cost
  - Not equivalent to buying a hard drive and putting it in your desk
- Need a sustained engagement with "cloud" computing?
- Also important for open data:
  Most research data services are not set up to provide unlimited / indefinite storage

From: Arthur Bullard
Sent: 11 December 2017 16:39
To: Jonathan Bright
Cc: Adham Tamer
Subject: Storage use on V drive

Hi Jonathan,

We are running low on storage space on the volume that serves the 'homes' (V drive) for all users, and noticed that your folder is 645GB which is over 10% of the total. If you are able to archive off some of this elsewhere, that would be very much appreciated, please. Or, you're very welcome to come and discuss alternative storage options with Adham or me.

Thanks!
-Arthur



Legal / ethical dimension

- Researchers can have easy access to highly sensitive material
  - Hate speech
  - People discussing illnesses, sexuality, political beliefs
  - Evidence of criminal activity
  - Often social media users will make mistakes / reveal more about themselves than they should
- No realistic possibility of obtaining “informed consent” in most cases
  - Mostly we work on “harm minimization”
- Site terms of service
  - Usually prohibits free sharing of data
  - Twitter would like you to store tweet IDs, not tweets (so people can delete it later)
  - Is it right to store immutable snapshots?
  - Should I publish raw data or processed datasets?

Massification

- We run a class where ~40 MSc students access large volumes of web data and write term papers on it. Many of them use the data for papers afterwards
- Research ethics committee can’t cope with 40 applications in a short turnaround time!
  - We have a “code of practice” instead for common cases
- Should all this data be saved and documented somewhere?

Conclusion

- Huge amount of opportunities to enhance social science with web streaming data
  - But also considerable challenges for the way current social science is run
- Some things I would like to see which enable this area more:
  - Short, small scale and rapid response grants -> e.g. ESRC open call is minimum £350,000 and takes 26 weeks to review. Can't respond quickly to things
  - Cautious approach to implementing open research data (e.g. ORD Concordat, RCUK principles). This is a good idea in theory but I feel enthusiasm is running ahead of practicality and assumes that there is huge value to be unlocked in all datasets -> this has caused problems for OD in the past!
    - Can't we share collection code rather than data?
  - More centralised IT provision (not sure if this is university or nationwide)
    - I want a large cloud file system which I access solely from my laptop. I want new accounts for each new student, the ability to share some files and keep others private, and then mark things for archival once I am finished / they underpin published research.

Thanks!

@jonmbright / jonathan.bright@oii.ox.ac.uk