Digital Data Infrastructures: interrogating the social media data pipeline. A talk originally given at the Association of Internet Researchers by Susan Halford, Mark Weal, Ramine Tinati, Les Carr & Catherine Pope, Web Science Institute, University of Southampton.

Introduction

WEB SCIENCE Institute

- From social media practices and effects … to data
- An unexpected gift bringing rich research opportunities
- Enthusiasm:

  '… it is as if the inner workings of private worlds have been pried open' (Latour 2007)
- Scepticism:

  '[w]hatever value big data may have for "knowing capitalism", its' value to social science has … [f]or the present at least, to remain very much open to question' (Goldthorpe 2016)
- A middle path in the space between *'giving in and getting out'* (Gehl 2015)




WEB SCIENCE Institute

Theorising data

- No such thing as ‘raw’ or ‘naturally occurring’ data
- All data are *‘always already social’* (Bowker 2013)
- We must explore:

  ‘… the lives and specificities of devices and data themselves, where and how they happen, who and what they are attached to and the relations they forge, how they get assembled, where they travel, their multiple arrangements and mobilizations and, of course their instabilities, durabilities and how they sometimes get disaggregated too’ (Ruppert, Law & Savage 2013)
- Where to start?

WEB SCIENCE Institute

The Data Pipeline

Database ⟷ Server Software ⟷ API(s) ⟷ Client Software ⟷ Subject

- Sociotechnical
- Iterative
- Core to the generation of data
- Core to the circulation of data
- Methodological implications?

WEB SCIENCE Institute

1: Population

- Demographics
- Location

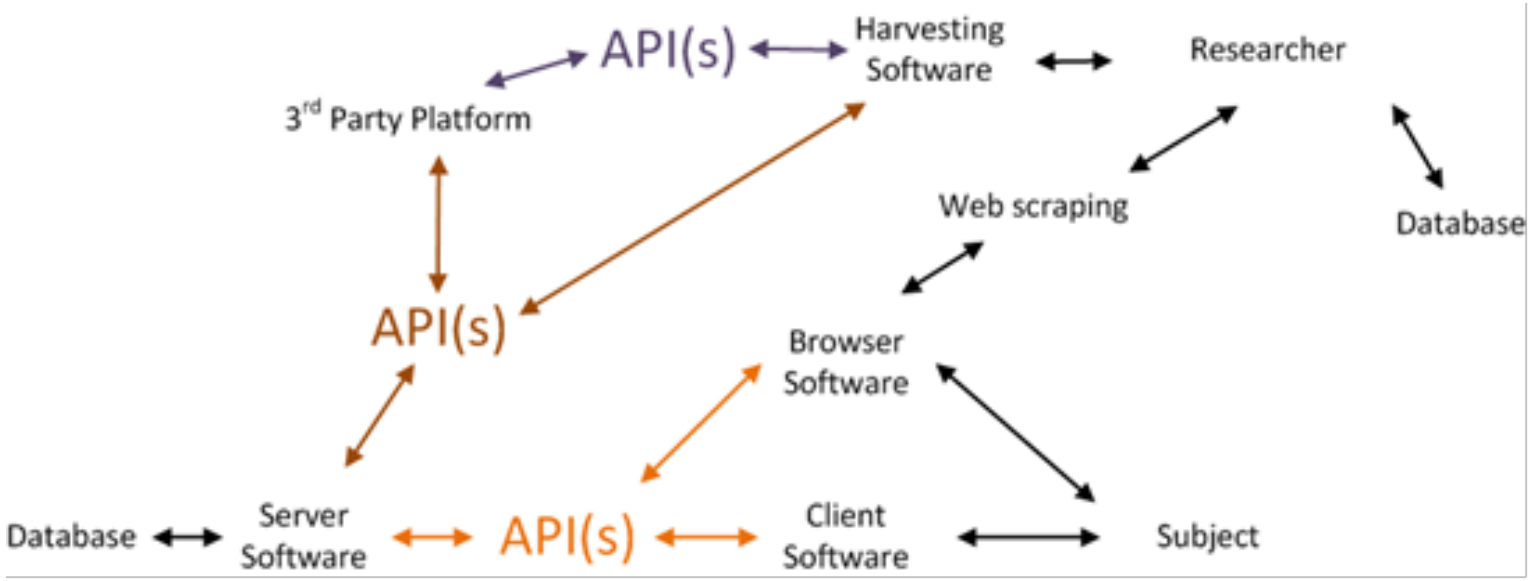
GPS location enabled - <3%
Jakarta 2.86%
Moscow 0.77%

1: Population

- Demographics
- Location
- Users – sovereign individuals?

Add corporate account image

2: Sample



- How the data are harvested shapes the sample
- The API shapes the sample e.g. % data streams, real time/historic
- Rate limiting

WEB SCIENCE Institute

3: Method

- Instruments for data collection
- Affordances
- For example: functionalities, data bases – shape data in specific ways over time

	Population	Sample	Method of data production
Database	Storage design and method shapes the types of information recorded about users.	Historic data storage decisions and technical query limitations may shape what data are included in samples.	Considerations of cost, performance and business requirements for data storage may shape what data are collected and stored and how.
Server Software	Determines who or what has access to the service, and what information is required to set up an account.	Server capacity may restrict data volume delivered; geographical location of server may affect data delivered.	Operates data management (e.g. spam removal and moderation, load balancing) shaping what data are collected.
API	APIs may not recognise all characters (languages) effectively; or be available to all operating systems/software development toolkits	A variety of differently structured samples may be available.	Defines the scope and volume of what data can be collected, stored and queried.
Harvesting Method	Harvesting methods construct different views of the populations. Web scraping may be more likely to access the population of currently active users, which could be different to the population accessed via historical searches using an API.	Web scraping will by-pass 'official' data samples, offering data from a sample of web pages. This sample may be affected by the 'filter bubble' of the person accessing the web pages. Use of third party data may introduce additional sampling effects.	Different harvesting methods have access to different types of data about the population and sample.
Client Software	Different clients may generate different information about the population. On some platforms you may know what client generated the content (this used to be the case on Twitter), on many though you can't know this.	Some clients (apps) may receive more data than others (if harvesting through a client).	Different clients may produce distinctive forms of data and metadata e.g. some may add geographic data by default, some might link directly to shared or re-shared material.
Subject	Different subjects – human/non-human, demographically distinct – may characterise particular platform populations.	User activities may shape sampling methods (e.g. official samples may focus on central or highly active users.)	User practices and meanings shape the data generated and the claims that can be made from these.

Conclusion

- Recognise the limits of what we can and can’t know about social media data

- Key steps

(1) Transparency
(2) Consider implications of data construction for research questions
(3) Knowledge claims
(4) Creative data assemblages