Selected Papers of AoIR 2016:
The 17th Annual Conference of the
Association of Internet Researchers
Berlin, Germany / 5-8 October 2016

DIGITAL DATA INFRASTRUCTURES: INTERROGATING THE SOCIAL MEDIA DATA PIPELINE

Susan Halford
Web Science Institute, University of Southampton, UK.
Mark Weal
Web Science Institute, University of Southampton, UK.
Ramine Tinati
Web Science Institute, University of Southampton, UK.
Les Carr
Web Science Institute, University of Southampton, UK.
Catherine Pope
Web Science Institute, University of Southampton,UK.

Social media are now firmly embedded across economies, cultures and politics and in the everyday lives of hundreds of millions of people, most famously a billion users signing on to Facebook in a single day. Particularly extraordinary is not just the new forms of social practice associated with social media – and their consequences – but that the very nature of these activities as digital and online constitutes them as a remarkable new source of social data. These data are generating widespread interest from business and government but the response from social scientists is mixed. On one hand, social media offers new insights to the things that people say and do 'in the wild', rather than the things they say they do in interviews and surveys, in real time, and at unprecedented a scale and pace (Savage and Burrows 2007, Burnap et al 2013, Tinati et al 2014; Weller et al 2013). On the other hand, some see these data as deeply problematic, flawed by demographic biases and unknown provenance. The well-established and well-regarded principles of social scientific research are grounded in clearly understood populations, carefully controlled sampling and well-known methods for collecting data, with a high value placed on transparency. Social media data offer none of this. Accordingly, it is suggested, this may lead not only to poor research and unsustainable claims but may ultimately risk the hard-won reputation of social science (Golthorpe 2016; Hardaker 2016).

This paper seeks a middle path in the space between *'giving in and getting out'* (Gehl 2015; 148): between accepting social media data at face value and abandoning the opportunities that they might offer. The key, we suggest, is methodological. Working with conventional sources of data, professional standards demand that we make the details of our research design, methods of data collection and data management

Halford, S., Weal, M., Tinati, R., Carr, L., Pope, C. (2016, October 5-8). *Digital Data Infrastructures: interrogating the social media data pipeline*. Paper presented at AoIR 2016: The 17th Annual Conference of the Association of Internet Researchers. Berlin, Germany: AoIR. Retrieved from http://spir.aoir.org.

explicit. This is rarely the case in publications that use social media data, which is typically presented with remarkably little methodological consideration of the data used. Certainly this is challenging. The most popular social media platforms are privately owned and make their data available, if at all, on their own terms and with differing levels of detail. However, just because social media are a novel, and opaque, source of secondary data is no less reason to consider these issues and their implications. To the contrary, there is all the more reason, if we are to allow social media data to be a credible and sustainable source for research.

We situate our investigation theoretically by drawing on the rich tradition of Science and Technology Studies, long used to conceptualise data infrastructures (Bowker and Starr 1999) and which informs recent theorisations of the broader 'dispositifs' (Ruppert et al 2012) or 'assemblages' (Kitchin and Lauriault 2013) that produce new forms of digital data. In short, whilst social media data have emerged from beyond the conventional practices of social science research, they are - despite the rhetoric sometimes deployed – anything but 'naturally occurring'. To the contrary, our theoretical approach insists that data are constructed through the actions of heterogeneous actors, from data bases, interfaces and browsers to consumers, markets and legal regulations.

On this wider landscape our specific focus in this paper is on the processes through which social media data are produced and made available to researchers. To investigate this we explore the 'pipeline' of social data production and circulation: from the user who creates the content, posting to a social media platform, to the client software on the phone, laptop, etc. that represents the data (sometimes in different ways, if there are multiple clients available for a given platform), the to the Application Programming Interface(s) (APIs), which enforces rules to determine what is passed through to the company's server software, and how, and the server software that organizes content into data bases that store data in particular formats and structures. This is a thoroughly sociotechnical process, shaped by technical interfaces and protocols, data storage and software applications. And by popular culture, business models, organizational resources and so on. In turn, all this shapes if and how these data are circulated for re-use, back down the pipeline. This 'output' is *not* a simple reversal of the 'input' and is shaped by the methods that researchers use to access data, the economics and practicalities for the companies in sharing data, with whom and on what basis, both shaped by legal and sometimes even ethical considerations.

Our paper explores how the data pipeline shapes three key methodological issues for researchers using social media data: the population, the sample and the method of data production. Taken together, this problematization of social media data may appear only to underscore the concerns expressed by those who have doubted their promise for robust social scientific research. This is not our intention. To the contrary, our tactic is to suggest that those of us using social media data should seek to address these challenges in our research. Certainly, we must accept that social media data are not like earlier generations of data, and consequently that the exact same methodological frameworks will not be appropriate. However, we should seek to position this new form of data methodologically, and develop new frameworks that will ensure its future value for researchers. We conclude by suggesting some methodological principles for the use of social media data that might strengthen – and thereby protect – this new source of

data for sociological research. Our conviction is that this will produce better academic research and will also develop our critical capacity to contribute to, and where necessary critique, the claims that are increasingly made from social media data by governments, the media and other commercial organizations.

Bowker, G., and Starr, S.L-, (1999) *Sorting Things Out: Classification and its' consequences* Cambridge, MA., MIT Press.

Burnap, P., Avis, N. and Rana, O. (2013) 'Making sense of self-reported socially significant data using computational methods' *International Journal of Social Research Methodology* 16(3), pp. 215-230**.**

Gehl, R. (2015) 'Critical reverse engineering' in Langlois, G., Redden, J., and Elmer, G. (Eds) C*ompromised Data: from social media to big data'* London, Bloomsbury.

Goldthorpe, J. (2016) *Sociology as a Population Science*, Cambridge, Cambridge UP

Hardaker, C. (2016) 'Misogyny, machines, and the media, or: how science should not be reported' http://wp.lancs.ac.uk/drclaireh/2016/05/27/misogyny-machines-and-the-media-or-how-science-should-not-be-reported/ Accessed 2 August 2016

Kitchin, R. and Lauriault, T. (2013) Towards Critical Data Studies: Charting and unpacking data assemblages and their work' *The Programmable City Working Paper 2*, University of Ireland Maynooth.

Ruppert, E., Law, J., and Savage, M. (2012) 'Reassembling Social Science Methods: the challenge of digital devices' *Theory, Culture and Society* 30(4) pp.22-46.

Savage, M., and Burrows, R. (2007) 'The coming crisis of empirical sociology' *Sociology*, 41(5) pp.885-899.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (Eds.) (2014) *Twitter and Society* New York, Peter Lang.

Tinati, R., Halford, S., Carr, L., & Pope, C. (2014) 'Big Data: Methodological Challenges and Approaches for Sociological Analysis' *Sociology* 48 (4) pp. 663-681.