First Monday, Volume 18, Number 10 - 7 October 2013

PEER-REVIEWED JOURNAL ON THE INTERNET

A critical reflection on Big Data: Considering APIs, researchers and tools as data makers

by Farida Vis

Abstract

This paper looks at how data is 'made', by whom and how. Rather than assuming data already exists 'out there', waiting to simply be recovered and turned into findings, the paper examines how data is co–produced through dynamic research intersections. A particular focus is the intersections between the application programming interface (API), the researcher collecting the data as well as the tools used to process it. In light of this, this paper offers three new ways to define and think about Big Data and proposes a series of practical suggestions for making data.

Contents

1. Introduction

In the summer of 2012, Twitter CEO Dick Costolo, during an interview on NBC's *Today* morning show, declared that he had figured out what the platform was all about, stating that: 'Twitter brings you closer' (Costolo in Wasserman, 2012). Later, at the Wired Business Conference he elaborated further, highlighting that there was a common shape to so many of the different stories on Twitter. Costolo emphasised how stories can be observed both from a distance, giving one perspective, as well as from a much closer, more personal vantage point, bringing you closer. To explain his point more elaborately, he went on to give a detailed description of the 2010 art installation in Tate Modern's Turbine Hall by the Chinese artist (and fêted Twitter celebrity) Ai Weiwei. Costolo describes the installation as follows:

> '[Weiwei] created this installation that was at the Tate museum in London a while back and the installation was these hundreds of thousands of ceramic hand painted sunflower seeds. And as you stood back from the room it looked like this sea of just stones that were black stones that were spread across the floor and of course you couldn't really tell what they were. But as you got closer it looks like ... you can start to tell 'ooh it looks like they've stamped out hundreds of thousands of sunflower seeds and spread them across the floor'. But as you pick them up you started to realise that they were all individually shaped and painted differently and unique and beautiful and distinct in their own right. *So that's what we want to bring to what we're building: the ability to shrink the world and allow everybody to see each other*.' (Emphasis added, Costolo in Vis, 2012a).

It can be assumed that by drawing on this catchy and accessible visual metaphor Costolo sought to render the Twittersphere both visible and comprehensible. What we can see, or rather imagine, by way of the art installation is the enormous volume of tweets that are produced daily on the platform. One billion tweets are now sent every two and a half days, making it hard data to deal with in retrospect, never mind in real time. Despite this volume, it seems Costolo encourages an understanding of Twitter that remains aware of the careful handcrafted and individual nature of tweets, sent by individual *people*. This is within the larger idea of the platform's ability to function as the proverbial global town square, an image the company has been keen to emphasise. By highlighting the art installation, the space within which the tweets exist, is also rendered visible to some extent, highlighting that this is a particular kind of space, a proprietary space. Costolo thus

gives us a glimpse of 'the world', offering up a number of different vantage points, not least his own, as a significant overseer of this world. Not just an overseer, but a creator also.

This paper is concerned with considering a number of ways in which data is made, by whom and how. Rather than assuming an *a priori* existence of the data 'out there', waiting to simply be recovered and turned into insight, this paper suggests that data is co–produced and that this production is premised on multiple dynamic research intersections. Specifically, this article is concerned with how data is made by application programming interfaces (APIs), looking at Twitter in particular. It considers how researchers themselves make and select data and finally, how the tools researchers use limit the possibilities of the data that can be seen and made. Tools then in turn take on a kind of data–making agency. In all of these data making processes and encounters, data is presented, selected and thought of in particular ways, with some data apparently considered more appealing for inclusion, processing and potential analysis than others. This means that some data becomes more visible than others. This tension between visibility and invisibility is nothing new. Yet the implications within the context of Big Data are worth considering anew, given the contemporary prominence and apparent popularity of this phenomenon.

The first part of the paper briefly considers a theoretical framework within which to think about Costolo's comments concerning visibility. It then highlights two influential definitions of Big Data, the first presenting an academic perspective, the second an industry one. The article seeks to further add to these from a critical academic position, offering three new ways to define and think about Big Data. It then considers why some online data have largely escaped the attention of Big Data's all–consuming gaze. With the strong focus on text and numerical data, and the revealing of networks and connections, images and other visual material uploaded on social media are rarely considered worthy of serious interest. This is peculiar given the evidential importance and value users themselves give to them. The article then moves on to discuss the ways in which APIs, researchers and tools each make data, concluding, and making some suggestions for further work.

2. Theoretical background and definitions

It is worth exploring this idea of a 'world' being created on Twitter and in turn how it can be made visible. This section draws on theoretical ideas of how states have traditionally sought to make different types of social worlds visible and in turn, controllable. It then considers two influential definitions of Big Data, compares them and suggests further additions.

2.1. Seeing a world (of data)

Twitter is of course not 'the world', it is 'a world', but what is of interest here is the way in which Costolo encourages people to think about Twitter as a particular world along with the vantage points available for seeing it. He therefore suggests two things: first, that it is indeed possible to see everything in this world, and second, that it is possible to make sense of what is seen. It is important to remember that what you see is framed by what you are able to see or indeed want to see from within a specific ideological framework. Not everyone sees the same things in this world. Costolo's statement chimes with some of the arguments James Scott (1998) has developed in *Seeing like a state* highlighting the power relations at work within certain visually controlled built environments. Scott details a series of failed attempts and grand schemes from the twentieth century, which shared the overall aim of seeking to improve the human condition. Driven by a belief in scientific laws, the concerted efforts on the part of these states was to make the lives of those within society more legible, therefore comprehensive and ultimately controllable by state powers through the application of a rational order. Scott emphasizes the role of airplanes in enabling a 'synoptic view' [1] along with the role of architecture and city planning, highlighting the redesign of Paris into a 'spectacle' [2], in order to make it easier to see everything and everyone. Noting the potentially detrimental experiential effects on the lives of citizens, he observes: 'The fact that such order works for municipal and state authorities in administering the city is no guarantee that it works for citizens' [3]. This statement is reminiscent of the increased debates around the experiences and rights of the users of social media platforms, the ones creating the data. Here the market is adopting and embracing similar ideals about synoptic views. Not designed to control in a governmental sense *per se*, though there is clearly a regime of control in place concerning the (re)use of data. Although the projects Scott describes ultimately failed, their enduring appeal has remained and it is perhaps unsurprising that many states now take a very keen interest in data collected from social media. Visibility can be instrumentalised in different ways, depending on the interests of those seeking to make something visible. Visibility can be useful as a means of control, it can be commercially exploited, or it can be sold to others who can exploit it in turn.

In describing both the distant and close up views, Costolo highlights something else besides: what can potentially be seen at a distance, 'just black stones', and what can be observed from close up: individually crafted aesthetically pleasing little sculptures. This means that whilst one might see, or thinks one sees, certain wider patterns from a distance, the view inevitable changes when focusing on objects up close. From up close these sunflower seeds are no longer seen as a homogenous mass of black stones, but are far more complex and messier to deal with. This is similar if we think of Twitter data or any smaller sample drawn from a much larger dataset for that matter. From a distance one billion tweets may look like a fairly homogenous mass, but from close up, at the level of the individual tweet, all sorts of messiness and complexity are revealed and it becomes very difficult to account for their specific context. This is difficult to deal with analytically, but that does not mean researchers should not try.

In highlighting the art gallery setting, Twitter is also revealed as a particular kind of space, where data is proprietary and for sale. Both aesthetically pleasing artefacts, sunflower seeds and tweets, possess monetary value as well. Two years after the show, the Tate announced that it had bought eight million of the porcelain sunflower seeds for an undisclosed sum, though a small number had sold for £3.50 a seed at Sotheby's the year before (Kennedy, 2012). This monetary value of the individual seeds resonates with the increased revenue generating initiatives at Twitter, through the selling of data by approved resellers such as Gnip and DataSift.

As various researchers have pointed out, this valuing of tweets and social data more widely in monetary terms, inevitably means data can become more unreliable due to distortion that happens across metrics. Metrics get gamed because money is involved. David Karpf (2012) has suggested that there is an urgent need for the online research community to accept that the data it draws on is 'likely never going to be all that good' [4]. He summarises the problem of gamed metrics as follows: *'Any metric of digital influence that becomes financially valuable, or is used to determine newsworthiness, will become increasingly unreliable over time'* [5]. What is then important is to be aware of how this specific unreliability can be accounted for in the research process. A number of practical suggestions are developed later on in this article. But before this, it is worth reflecting on how the concept of Big Data is defined and by whom.

2.2. Defining Big Data

Because of the broad interest in Big Data from a range of different groups across academia, government and industry, it is worth considering and comparing two widely cited definitions of Big Data arising from these different arenas, the first reflecting an academic perspective, the second an industry one.

The first definition, by boyd and Crawford (2012), offered in their recent influential article on the topic, includes both cultural and technological aspects, but also highlights Big Data as a 'scholarly phenomenon' [6]. They suggest that the interplay of these three specific aspects is crucial in defining Big Data. Quoting them in full, they propose the following three–part definition, resting on the interaction between:

1. *Technology*: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
2. *Analysis*: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
3. *Mythology*: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy. [7]

The emphasis on myth is important in two ways. First, because it draws on making visible the ways in which myths work and what is at stake in understanding this as a process. Roland Barthes (1993) explains that the key function of a myth is to naturalize beliefs that are contingent, making them invisible, and therefore beyond question (see Jensen, this issue). It is just so. boyd and Crawford argue that we are not yet at the stage where things are 'just so', but still in motion. Things are up for grabs so to speak, before the emerging ideas about Big Data become codified and institutionalised. There is therefore an urgent imperative to question the mechanisms and assumptions around Big Data. Related to this point is what Bowker and Star (2000) highlight about the limitations of available ways in which information can be stored in society. Instead of seeing the limitations of the technical affordances and imagine different ways in which information might be structured, the ways in which information is structured become naturalized, people begin to see these structures as 'inevitable' [8]. This last insight is useful in terms of developing critical ways for making them visible, for example, how social media companies make data available through APIs. What are the other possible ways in which data could be made available, thought about and imagined?

A second important definition of Big Data comes from industry, from IT consultancy company Gartner and essentially focuses on the first two parts of the first definition, emphasising technology and analysis. The Gartner definition of Big Data is as follows: '"Big data" is high–volume, –velocity and –variety information assets that demand cost–effective, innovative forms of information processing for enhanced insight and decision making' (Gartner in Sicular, 2013). Similar to the first definition, this is a definition in three parts. The first part consisting of the three V's — volume, variety and velocity — focused on the technical infrastructures necessary to deal with vast amounts of (unstructured) data. The next two parts highlight cost–effectiveness and innovation in processing this data. The final part focuses on the other key benefits: the possibility of greater insight and thus better decision–making. Unlike the first definition, this second one offers no possibility or encouragement to reflect on 'Big Data' as a phenomenon, or to make visible inherent claims about objectivity. Kate Crawford has recently extended the work around the mythology of Big Data to highlight the problem of '"data fundamentalism," the notion that correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth' (Crawford, 2013). This idea and belief in the existence of an objective 'truth', that something can be fully understood from a single perspective, again brings to light tensions about how the social world can be made known. Related to that: which methods and techniques (with their implied assumptions) can be used to know and understand the social world?

Returning to their definition and responding specifically to the focus of the second part of the boyd and Crawford's (2012) definition on analysis, as well as the emphasis on the three Vs in the second definition from Gartner, one might consider an alternative set of Vs from an academic perspective focused on the data–making process. They are listed below.

2.2.1. Validity

One of the key concerns within the social sciences in terms of using social media data is the anxiety around the validity and quality of the data. This takes a number of forms, both focused on the representativeness of the data. First, there is a focus on the representativeness of the data in relation to how the data is made available to the researcher. Second, there is a concern with how representative the data is in relation to a general off–line national population. So both are then concerned with how the sample is constructed and what can subsequently be read into this sample.

Again focusing on Twitter, a key concern that has recently produced some strong responses from academia, centres on better understanding how the Twitter firehose (all the publicly available tweets) relates to the various public APIs most academics have to rely on due to the cost involved in purchasing and processing firehose data. The key anxiety is not knowing exactly how Twitter samples from the firehose (although some documentation is publicly available) and this raises concerns about the overall value of this publicly available data (Gerlitz and Rieder; 2013; González–Bailón, *et al.*, 2012; Morstatter, *et al.*, 2013). This is thus a concern about making data at the level of Twitter creating the sample. Such issues around validity of the sample also arise in relation to the often opaque and unclear ways in which researchers themselves make and collect data for research purposes. The subsequent lack of sharing practices between academics in order for data to be examined by others and findings, in principle, becoming reproducible, remains an issue (Bruns, this issue).

A second key strand of concern highlights the lack of understanding of how Twitter users relate to general, national populations. In a recent study, led by Alan Mislove, *et al.* (2011), the researchers address this issue by examining over three million U.S. Twitter accounts, roughly one percent of the U.S. population and compare this to 2000 U.S. Census data. They examined the Twitter users along the axes of geography, gender, and race/ethnicity and conclude that Twitter users are significantly overrepresented in the more densely populated areas of the U.S., that users were predominantly male, and that they represented a highly non–random sample of the overall distribution of race and ethnicity. Other recent research by Eszter Hargittai and Eden Litt (2011), which highlights survey data from 505 diverse young American adults, shows that in their sample, young African Americans along with those with higher online skills are more likely to take up Twitter. They also suggest that interest in celebrity and entertainment was a significant predictor in Twitter use. In contrast, interest in news, local, national or international or politics showed no relationship to Twitter adoption in the population segment that they examined. Projects that seek to map national Twitter populations remain rare, with work led by key Twitter researchers Jean Burgess and Axel Bruns (2013), on mapping the Australian Twittersphere, a notable exception. In relation to the last study, though, it is worth introducing a final caveat. What is meant by a 'population'? Is this an online population in relation to an off–line one? So, is the question about how representative the online sample is in relation to what we know about the national population, derived for example from census data? Or is the question something else? In the case of the Burgess and Bruns study, this has a different starting point. This is about better understanding a population of Twitter users, as Twitter users. Drawing on the work of Richard Rogers (2009, 2013), this is about understanding online data as grounded in other *online data*, rather than off–line measurements. The online thus becomes the baseline. These are important fundamental differences to make explicit when we try to assess the multiple ways in which concerns arise over validity of samples, how these might be addressed, and which population they are meant to be related to exactly.

2.2.2. Venture

Thinking about the making process, 'venturing' can be a useful concept to explore and could mean a number of things: namely to offer (a take on the data), but the term can also hint at the more hazardous side of these practices. We venture into something, not presuming we already know all the answers. Thinking about venturing in this way also highlights the necessary curiosity required to want to find out about different possible worlds 'out there' expanding on what can be seen. This idea of venturing exists in contrast to frequent problematic assumptions that we can fully know the worlds we are investigating. Or, that this exploration is done from an objective position that is separate from the reality being discussed. Moreover, as Karpf (2012) highlights: 'The Internet is in a state of ongoing transformation' [9]. This then means that at best, we can venture to explore this highly dynamic and rapidly shifting world and offer partial glimpses across time. Finally, venturing is also concerned with offering a specific view when we present our findings, or a specific interpretation: We are on a mission to make a point about the data we made. This highlights the more overlooked, interpretative side of these practices, whether we deal with large or small volumes of data. As Costolo highlights in the opening anecdote, we tell stories about the data and essentially they are the stories we wish to tell.

2.2.3 Visibility

Issues around visibility have already been mentioned throughout this paper, but can be thought about in a further number of ways. Firstly, there is the visibility of the different steps taken in the process of making and dealing with data through a range of different encounters with the data. These traces often remain invisible, but can involve crucial information. Visibility also raises questions of invisibility and the tension between them already alluded to, namely what is and is not shown or is or is not seen. What is more, data is processed and turned into data visualisation: what does the data show? What does it not show and leave out? How can these visualisations be read and what kinds of skills are required for reading them? It is thus crucial to understand how visualisations are made, what they purport to show, what viewers think they show, and what they do not show. And finally, there is visibility in terms of online visual cultures, specifically in relation to the millions of images shared daily on social media, expanded on in the next section.

Following a brief provocation highlighting that not all data is of equal interest to the various research communities engaged with Big Data, the remainder of this paper is concerned with further exploring the analysis part of boyd and Crawford's (2012) definition, by looking at how data, prior to it being analysed is *made*, by whom and how, and what can be gained from understanding these processes better.

3. Provocation: Data not seen and not made

People currently produce and use a lot of images as part of their everyday lives. They do this primarily using digital media and online social media platforms and the rate at which they do this is rapidly increasing. In May 2011, 170 million tweets were sent daily, of which 1.25 percent contained a link to a picture from a photo sharing service indicating just over two million daily image shares (Levine, 2011). Similarly, two years after it was founded, in October 2010, Instagram reported that over 50 million people had shared over one billion photos through the app in October 2012 (Instagram, no date). Not one year later, in June 2013, Instagram boasted over 130 million active monthly users, 16 billion photos on the service altogether, with over one billion likes recorded *every single day* (Crook, 2013).

The use of online technologies to create and circulate images is increasing in popularity. Amongst U.K. Internet users, posting photos has significantly increased in recent years: from 44 percent of users in 2009 to 53 percent in 2011 [10]. These figures may be attributed to changes in mobile phone use in relation to which the taking and sending of photos are now the two most important activities after text messaging [11]. Moreover, within the Flickr photo–sharing community the Apple iPhone 4S is now the most popular camera used (Flickr, no date). These developments point to the increasing importance of smartphones — as all–in–one *networked* devices (Cruz and Meyer, 2012) — for the production and online circulation of personal photographs via platforms like Facebook and Twitter. The circulation of images on social media also involves the curation of 'personal digital collections' [12], often using images appropriated from elsewhere on the Internet, as indicated by the recent rise in popularity of the visual micro–blogging platform Tumblr and the image–collection platform Pinterest.

These practices involve the investment of considerable personal time and effort on the part of many people, suggesting that all sorts of cultural meaning and value are wrapped up in these activities, not least because social media allows people to show the images that they make or collect to others. This means that image related activities on social media are fundamentally communicatory and socially defined acts. As Van House and Davis (2005) point out, the presentation of personal photographs using mobile phones and social media has a number of 'social uses' that include the development and maintenance of relationships, the construction of personal and collective memory, self–presentation, and self–expression. Such social circulations of images involve collectively shared values. But more than that, they represent meeting points between personal and collective value. People send, collect, organise, and show images of people (and therefore relationships), objects, and experiences they value. They also send, collect, organise, and show already existing images that they value, often because they are valued and have meaning in a wider collective cultural context (Popescu, 2012; Zarro and Hall, 2012). Such images can be valued as forms of witnessing, as aesthetic artefacts, documents, historical records, as tokens of collective identity, and so on.

Social media companies value images differently. One could argue for example that Facebook values images in terms of their ability to attract certain types of activities by the users of the platform. This occurs most notably through enriching the image with tags, so that connections can be gleaned between users: they both appear in the same image (which may also include additional data such as location). As of January 2013 Facebook had 240 billion images stored on its service (Wilhelm, 2013). With many users tagging and enriching this data with what is likely to be reliable information (provided the tags are related to an identifiable user and also correctly identifies this user). This thus means that this large dataset provides important opportunities for machine learning and training algorithms.

Yet images are currently an under–researched area within social media research (Vis, *et al.*, 2013) as well as within the growing interest in Big Data. The images themselves do not easily lend themselves to popular Big Data 'mining' techniques and are thus typically a discarded data object in such enquiries. It therefore seems that in an academic research context at least, images are not as valued as they are by social media users themselves. The question for researchers then becomes how images can or should be valued within a research context, seen as valuable research objects within Internet research. Some suggestions about how this might be done analytically are offered in the next section. What is clear, however, is that due to the complex nature of current production, viewing (think Snapchat for example) and circulation practices, we need to identify and draw on a range of different relevant theories and methods to make sense of these emerging visual cultures on social media.

4. Practices: APIs, researchers and tools making data

Focusing now on how we can think about Big Data analysis as a practice is important. Karpf (2012) emphasises that Internet researchers should be 'question–driven, letting the nature of their inquiry determine their methods' [13]. This seems an obvious point, but in practice it is not always in evidence in current social media research. Sometimes research arises simply because data is available (for example through the donation of data), which then greatly limits the questions than can be asked because data was not originally created with the questions we now wish to ask, in mind. Data made available through APIs, including paying attention to how this is made available, can further limit the questions researchers ask. Moreover, the tools we use can limit the range of questions that might be imagined, simply because they do not fit the affordances of the tool. Not many researchers themselves have the ability or access to other researchers who

can build the required tools in line with any preferred enquiry. This then introduces serious limitations in terms of the scope of research that can be done.

4.1. APIs making data

Most social media platforms now make data publicly available on their platform through APIs. This has resulted in considerable development from third party developers to create all sorts of applications and content on top of this data. Two key companies that have emerged as key data brokers, specifically focusing on social media data from a large number of sources, are Gnip and Datasift. They have either been established recently or have refocused their attention on social media data. Gnip was founded in 2008, but since 2010 has had a strong focus on social media data. Social media platforms do not give access to 'all' the data, though. For example, data derived from enhancing activities will not necessarily be made available for sale. Added value, created for instance when a link is shortened on Twitter and turned into 'a t.co link', provides additional data, which is currently not shared. The t.co link is the standard way in which Twitter shortens links: it 'measures information such as how many times a link has been clicked, which is an important quality signal in determining how relevant and interesting each Tweet is when compared to similar Tweets' (Twitter, no date). At the time of writing, Gnip highlights that it gives full firehose access to Twitter, FourSquare and Tumblr data, while public API data is included for Facebook, Flickr, YouTube, Instagram, Vimeo, Google+ and others. This in principle means that tools could be developed that draw on this rich API data ecology and pull in data from a range of different platforms to allow for comparative analysis. While the technical data infrastructure has reached a point where this is possible, academia has yet to catch up with these possibilities. Although industry 'listening platforms' do tend to pull in data from a range of different platforms to track conversations, these tools are often unsuitable for academic purposes because of their cost, along with the problematic 'black box' nature of many of these tools.

Already extensive metadata is now regularly 'enriched' further by companies like Gnip and Datasift to produce rich opportunities for advertisers to mine it. Given the data explosion associated with social media data, it is then not surprising that an emerging industry has developed around so called 'social data', largely focused on metadata.

Gnip is keen to highlight the difference between 'social media' and 'social data'. For Gnip social media is essentially about communication and users expressing themselves, where their content is 'delivered to other users'. Social data on the other hand 'expresses social media in a computer–readable format (*e.g.*, JSON) and shares metadata about the content to help provide not only content, but context. Metadata often includes information about location, engagement and links shared. Unlike social media, social data is focused strictly on publicly shared experiences' (Ellis, 2013). As Bruhn Jensen (this issue) highlights about metadata, this data about data is a source of information in its own right, beyond what is delivered, 'sent', to other users and in turn 'received' by them. This 'meta–information' seemingly situates the content exchanged through this communication in relation to its contexts. Understanding how social media platforms make metadata along with what they make available is in constant flux. More than that, we need to better understand how the metadata offered for sale or through public APIs is often 'enriched' in ways that gloss over a variety of problems associated with this process.

For example Gnip has recently started offering something it calls 'Profile Geo enrichment' (Cairns, 2013), in order to help a 'customer understand offline locations that are relevant to online conversations' (Cairns, 2013). Location data is highly sought after within industry and academia. Gnip's company blog highlights that this new product is able to 'normalize' all the different ways in which people leave location traces online, in the first instance focusing on Twitter, stating that their customers are 'hungry to analyze Twitter through a geographic lens' (Cairns, 2013). The blog also explains how this new product relates to data available through the Twitter APIs, including the firehose. Attention is drawn to the fact that whilst only one percent of users engage in geolocation activities (resulting in fewer than two percent of all tweets containing latitude/longitude coordinates), more than half of tweets contain a location in the profile information of the user. For this reason, the profile location is thus seen as highly relevant, not least because, as Gnip argues, this offers more 'evenly distributed' information (presumably to mean 'more representative') across the Twittersphere. In other words, Gnip does not simply focus on this fraction of users that add latitude/longitude coordinates to their tweets. A product is thus created around making 15 times more geolocation data available than currently possible on Twitter, so that Gnip customers 'can now hear from the whole world of Twitter users and not just this 1 percent' (Cairns, 2013). Finally, making a leap beyond Twitter in line with Mislove, *et al.'s* (2011) research into connecting Twitter data to the general U.S. population, Gnip also emphasizes similar opportunities for grounding profile location data more widely in government–derived datasets:

> Profile location data can be used to *unlock* demographic data and other information that is not otherwise possible with activity location. For instance, U.S. Census Bureau statistics are aggregated at the locality level and can provide basic stats like household income. Profile location is also a strong indicator of activity location when one isn't provided. (Emphasis added; Cairns, 2013)

Going back to the idea of a synoptic view, what we can see here is the combination of what Costolo describes in terms of seeing the world through Twitter and, theoretically, through the techniques outlined by Scott (1998), which enabled governments to see society. In this case however, the ones seeing are commercial companies that sell the data to advertisers, keen to more accurately target specific groups of (potential) customers. Whilst the data company acknowledges the limitations of these enrichment methods, these will be hard to trace once this enhanced metadata is situated within a large database and layer on layer on metadata has been added to a user. This is especially the case when this data is also combined with additional

metadata, not derived from Twitter (so pertaining to another data infrastructure), as well as further enrichment through government statistics as outlined above. It seems evident how difficult it then becomes to assess and unpick this data once it appears in aggregate.

In early 2013 it was reported that Twitter itself was also releasing richer metadata through its various APIs. Following a series of recent restrictions within the Twitter API data ecology, self–described educational tech explorer, Martin Hawksey, highlighted in March 2013 that the Twitter Search API now gave access to considerably more metadata than before (Hawksey, 2013). Whilst there were many similar blog posts by developers at the time, including comments on the main announcement on the official Twitter developers blog, Hawskey does something very interesting: he *shows* how much more data can now be accessed through the Twitter Search API (see Figure 1). This gives an instant overview in terms of the massive expansion in data created and in turn made available through the APIs.

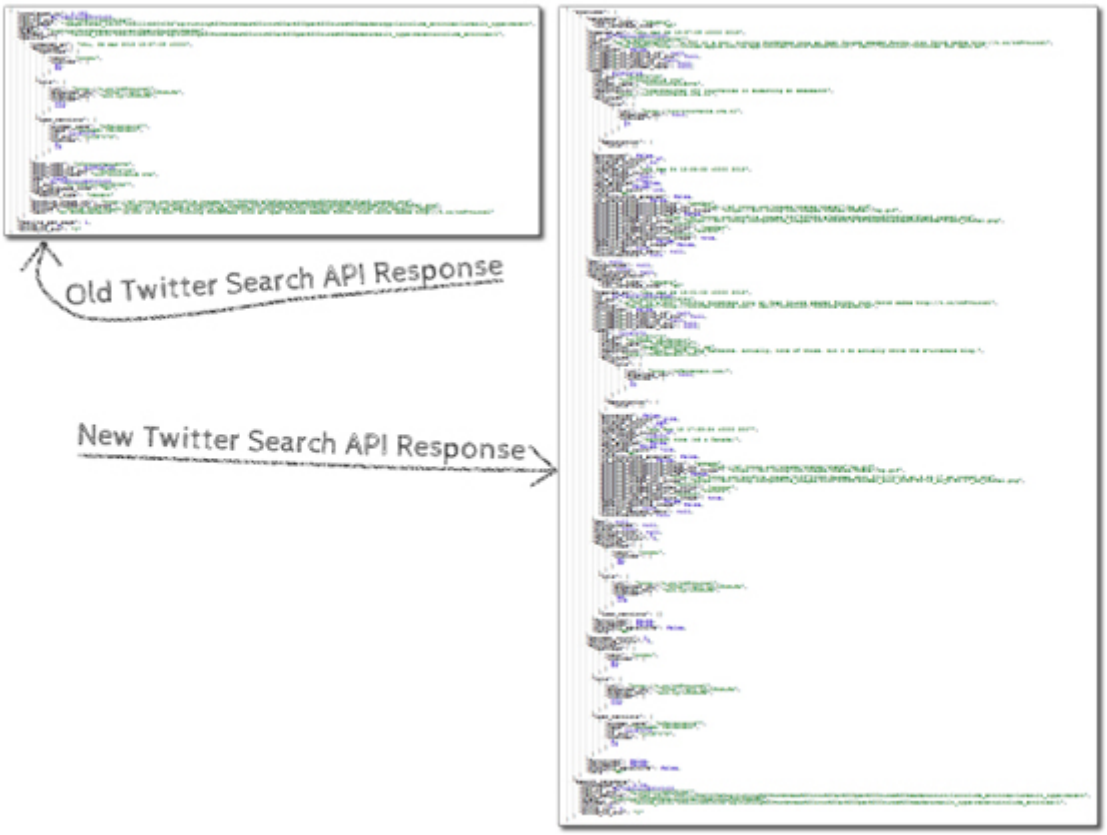

Figure 1: Old and new Twitter search API responses (March 2013).

Comparing the image on the left, the old API data, to the one the right, it is instantly clear that the new version of the Twitter Search API gives access to far more metadata than before. For the purposes of this paper it is worth considering whether this constitutes additional access to metadata that already existed but developers only now received access to, *or* if this additional metadata has been created, *made* by the company in order to attract more development activity around the platform. Put differently, the API itself could be seen as a data maker here. Not in the way described earlier, highlighting anxieties over the way data is sampled from the firehose, but rather as a dynamic service that constructs and gives access to new and 'enriched' metadata, pandering to the perceived desires of a seemingly insatiable data market. Writing about the politics of Twitter data, Puschmann and Burgess (2013) note the lack of consideration or visibility of the end user in these data making and trading activities: 'End–users (both private individuals and non–profit institutions) are without a place in it, except in the role of passive producers of data. The situation is likely to stay in flux, as Twitter must at once satisfy the interests of data traders and end–users, especially with regards to privacy regulation. However, as neither the contractual nor the technical regulatory instruments used by Twitter currently work in favour of end users, it is likely that they will continue to be confined to a passive role'. This statement resonates with Scott's (1998) observation about the potentially detrimental experience for citizens as part of the government structures he outlines.

4.2. Researchers making data

Researchers of course also actively make data, though this can easily be overlooked or not attributed enough consideration. How a dataset is made, including which search strategies were employed, are rarely included in research findings. Whilst there is evidently no room for lengthy discussions of this kind in a standard journal article, it is worth recording this information either in an appendix or elsewhere. Decisions about what to collect (what is in, what is out), from which API data is collected, for which period, including which metadata, including an awareness of how this collected data is itself created by APIs, are important stages in the data making process. Researchers should aim to make themselves more aware and reflect more on the process through which they have collected data and make this as transparent as possible. In line with Karpf's (2012) similar call for greater transparency, as researchers we 'should be up–front about the limitations of our data sets and research designs' [14]. Expanding on that call, it is also important to know what the limitations are in the first place. Given the descriptions of how APIs create data, this is not always a straightforward process to engage with even if the intention is to be as transparent as possible in the research process. As a community of researchers, we also have to become better at describing some of the overarching problems in *principle*, but also in *practice* seeking to explain what this actually entails, using examples where necessary.

The suggestions presented below are premised on the author's own experience of working with a mixed methods approach, where the collection of an initial Twitter dataset is typically followed by a much closer examination of a subset of this larger dataset. The focus of this small subset can be multiple, for example the examination of individual journalists and the content they have created (Vis, 2012b) or a select set of images shared during a specific event (Vis, *et al.*, 2013). Among Twitter researchers within and outside academia, it has for some time now been clear that there is need to move away from often quite simple data making strategies. In light of this, it is worth thinking about the techniques that may be useful in making better data and which disciplines may help us here.

4.2.1. More sophisticated hashtag/keyword search strategies

The research area of information retrieval (IR) offers some valuable possibilities for developing more comprehensive search strategies. In building a corpus, IR would typically treat this activity as a multi–staged process with a number of stages that build on previous ones by suggesting further informed search strategies [15].

Starting with one set of search terms, this would produce an initial set of tweets. New hashtags and keywords will be present in this data and co–location techniques could produce a new set of terms to subsequently search for. In principle you could do this until no further useful search terms are identified, though it may be difficult to assess when that point is reached. When do you know a tweet is still about your topic? This is highly context dependent and difficult to gauge at a distance, by simply seeing keyword/hashtags offered for consideration by a research tool for example. Moreover, it is possible that in some cases the same term can both be about your topic and also not be about your topic depending on the context. This then requires manual checking and domain expertise to resolve and a description of how this was resolved in the dataset.

4.2.2. Search based on links

Once these search techniques have been exhausted, more relevant data may be collected by searching again, but this time for frequently shared links [15]. These could be popular newspapers articles, blog posts, videos, or images present in the data. By applying this extra search approach one — to some extent also — gets beyond the problem of language bias that may otherwise be an issue. It is often difficult to build a strong set of candidate terms in other languages. This second technique may then identify additional data, for example tweets that use none of the hashtags or keywords, but simply share a link with a comment (for example: 'I can't believe this is happening right now': link)

Users can experience images on Twitter in one of two key ways: by accessing and viewing the image directly on the platform (if uploaded directly to Twitter) or by clicking a link that takes you to the image, for example on Instagram. At the time of writing, most are uploaded external to Twitter and shared on the platform. For researchers, it is thus challenging to deal with images that for social media users are available in accessible visual forms, but as data are effectively invisible, because they entail machine–readable hyperlinks. Making these images visible again is not easily done at scale. Researchers therefore need to find ways to navigate and gain visual access to smaller groups of images within large datasets. Consequently the analytical journey from macro–scale Big datasets to micro–scale small data interpretation is a complex one, but could fruitfully include methodological experiments in how to access images *through* the data. Once images are accessible, interpretative approaches appropriate to the specificity of images circulated through Twitter are required. These approaches will need to be sensitive to differences between images in terms of medium, genre, aesthetic form, and types of spectatorship. Traditional quantitative approaches are also not well equipped to give such insights.

4.2.3. Duplicate detection

Once the initial dataset has been made, further inspection is often needed to address two additional key issues: duplicate content and spam. Especially when collecting large volumes of data, filtering out spam content may be required. A combination of human verification and computational filtering can be very productive. What constitutes 'spam' for the purposes of the specific project may also need to be considered and is likely to be project specific.

It is also important to explain how duplicate content was dealt with. Duplicate content in the context of Twitter most notably means how retweets, modified tweets and automated retweets are identified. Manual retweets are easily identified by searching for the letters 'RT', whilst the identification of automated tweets requires an algorithmic intervention, which instructs an algorithm to compare tweets to each other and

identify identical or near identical content (for a description of this technique, see Lotan, *et al.*, 2011). What constitutes 'near identical' content will require further explanation, namely by highlighting how the algorithm operates in terms of the Levenshtein distance used to essentially say 'tweet b is still similar enough to tweet a because the modification resides within our agreed set parameters' (for a straightforward explanation, see Gilleland, no date). As tweets are such short pieces of text, a seemingly small moderation, for example the adding of '+1!' to an original tweet can quite drastically alter its meaning. It is thus imperative to explain how such potential duplicate content was dealt with in order to better understand the findings.

Finally, and this is incredibly hard to deal with at a large scale, it is important not to overlook the issue of potential problems of data including the effects of gamed metrics, including click farms, fake/spoof profiles, follower boosts, bots, and other forms of deception like swarms, link–baiting and so on, already highlighted. This then highlights an important aspect of dealing with Twitter data: it important to have domain expertise, as this will facilitate the ability to recognise limitations or identify anomalies in a given dataset. It is therefore productive to distinguish between Big Data approaches that treat Twitter data simply as a lot of data and those that approach Twitter as social media first, often as part of a media/cultural/Internet studies approach.

4.2.4. Dealing with feature changes

It is also important to be aware of the dynamic nature of the Twitter data ecosystem and platform specificities that may need to be considered as part of the data collection strategy. Twitter data collected in 2009 is difficult to compare to data collected in 2013. Features are created, updated and retired, and it is thus important to take into account the implications that may arise for making and comparing data. If such feature changes occur during the collection of data, it important to make these explicit and discuss the potential implications that may arise across the collected data.

4.3. Tools making data

Due to the analytical challenges of Big Data it is unsurprising that researchers working in this area now regularly operate as teams or research groups, and that they will now most likely include a computer scientist who can aid with the data collection and processing. It is worth noting that within such setups the making of the data is often not done by the whole team, and it thus becomes important to reflect on such strategies. What are the benefits of a team rather than a single person discussing the collection strategy and actually gathering the data? This highlights an issue about skills within such interdisciplinary teams: who is actually capable of collecting the data in the first place? Moreover, such teams will often build a set of bespoke tool solutions that others cannot use. Or, even if tools built by other computer scientists are freely available, it remains commonplace for a newly formed team to build its own tools. As a consequence of such practices tools for researching Twitter have developed in silos, and where tools are freely available it is not clear how widely these are used and what the barriers to their adoption may be. Although there have been recent calls to aim to standardise approaches within Twitter research (Bruns and Stieglitz, 2013), little attention has yet been paid to important role tools themselves play in terms of data making. Unlike industry, where different types of 'listening platforms' and tools are frequently reviewed, academia has been slow to offer similar reviews of the different tools now available for studying Twitter. Aside from the occasional discussion about the problems of the 'black box' nature of many tools, it would be incredibly helpful to critically consider what we want from our tools. What are our requirements as researchers and how can these be translated into future tool development? That is to say: what are our research questions?

Two tools created by information scientist Mike Thelwall for example, Webometric Analyst (http://lexiurl.wlv.ac.uk/) and Mozdeh (http://mozdeh.wlv.ac.uk/) are worth considering in this context. Webometric Analyst allows for the collection of social media data, including from Twitter, and offers a range of methodological approaches, such as network analysis and sentiment analysis. Mozdeh focuses on longitudinal data collection from Twitter and encourages time–series analyses, allowing the researcher to observe changes over time. The fact that these newer generations of tools move far beyond the simple archival attributes of now defunct services like TwapperKeeper is worth noting. Within their design, more critical approaches to data collection and analysis are embedded from the start. This is particularly true of Mozdeh. As tools evolve further it is important to understand and question the research limitations that are implicit in their design. Do they stop researchers asking certain types of questions? If so, what can we do about this?

Alongside a more critical understanding of the role of APIs in creating data, reflecting on our own roles as researchers in making data and describing its limitations, we also need to cast a critical eye on the tools we use. In doing this, we need to move beyond discussions about their perceived 'black box' nature, although these remain important. We need to expand our research imaginations and start first and foremost with identifying the research questions we wish to ask. Seeing through and accounting for the ideological assumptions at the heart of Big Data (boyd and Crawford, 2012; Crawford, 2013), requires more than a response that highlights the importance of showing the limitations of the questions that can be asked of Big Data. Whilst this is incredibly important work, it is also critical to have discussions about the questions we wish to ask and how we might answer them, including a careful consideration of the data, methods and tools we might need to do so. Designed with the research question as the dominant driver of the enquiry, such approaches could further develop the area of social media research in exciting new ways in the future.

About the author

Farida Vis is a Research Fellow looking at 'Big Data and Social Change', based in the Information School at the University of Sheffield, United Kingdom.

Twitter: @flygirltwo
E–mail: F [dot] vis [at] Sheffield [dot] ac [dot] uk

Acknowledgements

The author is grateful to the editors of this special issue, the reviewers and Simon Faulkner, who have all made invaluable comments on earlier drafts of this paper.

Notes

1. Scott, 1998, p. 58.

2. Scott, 1998, p. 62.

3. Scott, 1998, p. 58.

4. Karpf, 2012, p. 649.

5. Karpf, 2012, p. 650.

6. boyd and Crawford, 2012, p. 663.

7. boyd and Crawford, 2012, p. 663.

8. Bowker and Star, 2000, p. 108.

9. Karpf, 2012, p. 647.

10. Dutton and Blank, 2011, p. 29.

11. Dutton and Blank, 2011, p. 15.

12. Feinberg, *et al.*, 2012, p. 200.

13. Karpf, 2012, p. 641.

14. Karpf, 2012, p. 652.

15. I am grateful to Paul Clough and Mike Thelwall for sharing their insights with me.

16. I thank Francesco D'Orazio for extensively discussing these strategies with me recently.

References

Roland Barthes, 1993. *Mythologies*. Selected and translated by Annette Lavers. London: Vintage.

Geoffrey C. Bowker and Susan Leigh Star, 2000. Sorting things out: Classification and its consequences. Cambridge, Mass.: MIT Press.

danah boyd and Kate Crawford, 2012. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society*, volume 15, number 5, pp. 662–679.
doi: http://dx.doi.org/10.1080/1369118X.2012.678878, accessed 24 September 2013.

Axel Bruns and Stefan Stieglitz, 2013. "Towards more systematic Twitter analysis: Metrics for tweeting activities," *International Journal of Social Research Methodology*, volume 16, number 2, pp. 91–108.
doi: http://dx.doi.org/10.1080/13645579.2012.756095, accessed 24 September 2013.

Jean Burgess and Axel Bruns, 2013. "Mapping the Australian Twittershere," paper presented at Media in Transition 8 Conference (3–5 May, Boston); slides at http://www.slideshare.net/Snurb/mit8-burgessbruns, accessed 24 September 2013.

Ian Cairns, 2013. "Get more geodata From Gnip with our new profile geo enrichment," Gnip Company Blog (22 August), at http://blog.gnip.com/tag/geolocation/, accessed 13 September 2013.

Kate Crawford, 2013. "The hidden biases in Big Data," *Harvard Business Review (HBR) Blog Network* (1 April), at http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/, accessed 10 September 2013.

Jordan Crook, 2013. "Instagram crosses 130 Million Users, with 16 billion photos and over 1 billion likes per day," *TechCrunch* (20 June), at http://techcrunch.com/2013/06/20/instagram-crosses-130-million-users-with-16-billion-photos-and-over-1-billion-likes-per-day/, accessed 4 September 2013.

Elaine Ellis, 2013. "Social data vs social media," *Gnip Company Blog* (3 May), at http://blog.gnip.com/social-data-vs-social-media-2/, accessed 10 September 2013.

Edgar Gómez Cruz and Eric T. Meyer, 2012. "Creation and control in the photographic process: iPhones and the emerging fifth moment of photography," *Photographies*, volume 5, number 2, pp. 203–221. doi: http://dx.doi.org/10.1080/17540763.2012.702123, accessed 24 September 2013.

William H. Dutton and Grant Blank, 2011. "Next generation users: The Internet in Britain," *Oxford Internet Survey 2011 Report*, at http://www.worldinternetproject.net/_files/_Published/23/820_oxis2011_report.pdf, accessed 18 August 2013.

Melanie Feinberg, Gary Geisler, Eryn Whitworth, and Emily Clark, 2012. "Understanding personal digital collections: An interdisciplinary exploration," *DIS '12: Proceedings of the Designing Interactive Systems Conference*, pp. 200–209. doi: http://dx.doi.org/10.1145/2317956.2317988, accessed 24 September 2013.

Flickr, no date. "Camera finder," at http://www.flickr.com/cameras, accessed 15 April 2013.

Carolin Gerlitz and Bernhard Rieder, 2013. "Mining one percent of Twitter: Collections, baselines, sampling,&edquo; *M/C Journal*, volume 16, number 2, at http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620, accessed 14 September 2013.

Michael Gilleland, no date. "Levenshtein distance, in three flavors," at http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm, accessed 10 September 2013.

Sandra González–Bailón, Ning Wang, Alejandro Rivero, Javier Borge–Holthoefer, and Jamir Moreno, 2012. "Assessing the bias in communication networks samples from Twitter," at http://arxiv.org/ftp/arxiv/papers/1212/1212.1684.pdf, accessed 14 September 2013.

Eszter Hargittai and Eden Litt, 2011. "The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults," *New Media & Society*, volume 13, number 5, pp. 824–842. doi: http://dx.doi.org/10.1177/1461444811405805, accessed 24 September 2013.

Martin Hawksey, 2013. "Twitter throws a bone: Increased hits and metadata in Twitter Search API 1.1" (28 March), at http://mashe.hawksey.info/2013/03/twitter-throws-a-bone-increased-hits-and-metadata-in-twitter-search-api-1-1/, accessed 10 September 2013.

Instagram, no date. "2 years later: The first Instagram photo," at http://blog.instagram.com/post/27359237977/2-years-later-the-first-instagram-photo, accessed 4 September 2013.

David Karpf, 2012. "Social science research methods in Internet time," *Information, Communication & Society*, volume 15, number 5, pp. 636–661. doi: http://dx.doi.org/10.1080/1369118X.2012.665468, accessed 24 September 2013.

Maev Kennedy, 2012. "Tate buys eight million Ai Weiwei sunflower seeds," *Guardian* (5 March), at http://www.theguardian.com/artanddesign/2012/mar/05/tate-ai-weiwei-sunflower-seeds, accessed 14 September 2013.

Sheldon Levine, 2011. "How people currently share pictures on Twitter," *Sysomos Blog* (2 June), at http://blog.sysomos.com/2011/06/02/, accessed 14 September 2013.

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd, 2011. "The revolutions were tweeted: Information flows During the 2011 Tunisian and Egyptian revolutions," *International Journal of Communication*, volume 5, at http://ijoc.org/index.php/ijoc/article/view/1246, accessed 24 September 2013.

Alan Mislove, Sune Lehman, Yong–Yeol Ahn, Jukka–Pekka Onnela and J. Niels Rosenquist, 2011. "Understanding the demographics of Twitter users," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 554–557, at https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816, accessed 14 September 2013.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley, 2013. "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose," *Association for the Advancement of Artificial Intelligence Conference*, at http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf, accessed 14 September 2013.

Richard Rogers, 2013. *Digital methods*. Cambridge, Mass.: MIT Press.

Richard Rogers, 2009. *The end of the virtual: Digital methods*. *Oratiereeks/University of Amsterdam, Faculty of Humanities*, number 339. Amsterdam: Vossiuspers UvA.

Ana–Maria Popescu, 2012. "Pinteresting: Towards a better understanding of user interests," *DUBMMSM '12: Proceedings of the 2012 Workshop on Data–Driven User Behavioral Modelling and Mining From Social Media*, pp. 11–12. doi: http://dx.doi.org/10.1145/2390131.2390136, accessed 24 September 2013.

Cornelius Puschmann and Jean Burgess, 2013. "The politics of Twitter data," In: Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (editors). *Twitter and society*. New York: Peter Lang.

John C. Scott, 1998. *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven, Conn.: Yale University Press.

Svetlana Sicular, 2013. "Gartner's Big Data definition consists of three parts, not to be confused with three 'V's," *Forbes* (27 March), at http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/, accessed 18 August 2013.

Twitter, no date. "About Twitter's link service (http://t.co)," *Twitter Help Center*, at https://support.twitter.com/articles/109623-about-twitter-s-link-service-http-t-co, accessed 15 September 2013.

Nancy A. Van House and Marc Davis, 2005. "The social life of cameraphone images," *Proceedings of the Pervasive Image Capture and Sharing: New Social Practices and Implications for Technology Workshop (PICS 2005) at the Seventh International Conference on Ubiquitous Computing (UbiComp 2005)*, at http://people.ischool.berkeley.edu/~vanhouse/Van%20House,%20Davis%20-%20The%20Social%20Life%20of%20Cameraphone%20Images.pdf, accessed 14 September 2013.

Farida Vis, 2012a. "'Twitter brings you closer': The importance of seeing the little data in Big Data," In: Drew Hemment and Charlie Gere (editors). *FutureEverybody: FutureEverything Report*, pp. 43–45, at http://futureeverything.org/FutureEverybody.pdf, accessed 10 September 2013.

Farida Vis, 2012b. "Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 UK riots," *Digital Journalism*, volume 1, number 1, pp. 27–47.
doi: http://dx.doi.org/10.1080/21670811.2012.741316, accessed 24 September 2013.

Farida Vis, Simon Faulkner, Katy Parry, Yana Manykhina, and Lisa Evans, 2013. "Twitpic–ing the riots: Analysing images shared on Twitter during the 2011 UK riots," In: Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (editors). *Twitter and society*. New York: Peter Lang.

Todd Wasserman, 2012. "Costolo on Twitter's purpose: 'We bring people closer'," *Mashable* (18 September), at http://mashable.com/2012/09/18/costolo-today-show/, accessed 18 August 2013.

Alex Wilhelm, 2013. "Facebook: Our 1 billion users have uploaded 240 billion photos, made 1 trillion connections," *TNW* (15 January), at http://thenextweb.com/facebook/2013/01/15/facebook-our-1-billion-users-have-uploaded-240-billion-photos-made-1-trillion-connections/, accessed 13 September 2013.

Michael Zarro and Catherine Hall, 2012. "Pinterest: Social collecting for #linking #using #sharing," *JCDL '12: Proceedings of the 12th ACM/IEEE–CS Joint Conference on Digital Libraries*, pp. 417–418.
doi: http://dx.doi.org/10.1145/2232817.2232919, accessed 24 September 2013.

Editorial history

Received 16 September 2013; accepted 17 September 2013.

A critical reflection on Big Data: Considering APIs, researchers and tools as data makers
by Farida Vis.
First Monday, Volume 18, Number 10 - 7 October 2013
http://journals.uic.edu/ojs/index.php/fm/rt/printerFriendly/4878/3755
doi:10.5210/fm.v18i10.4878.