Alexander von Humboldt
INSTITUT FÜR INTERNET
UND GESELLSCHAFT

HIIG Discussion Paper Series

Discussion Paper No. 2013-01

The Politics of Twitter Data[1]

23. Jan. 13

Cornelius Puschmann
cornelius.puschmann@oii.ox.ac.uk
Oxford Internet Institute (OII)
University of Oxford

Jean Burgess
je.burgess@qut.edu.au
ARC Centre of Excellence for Creative Industries and Innovation
(CCI)
Queensland University of Technology

[1] This paper is a draft chapter from the forthcoming volume *Twitter and Society* (K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann, eds.) which will be available from Peter Lang Publishers, NYC, in spring 2013.

Abstract

Our paper approaches Twitter through the lens of "platform politics" (Gillespie, 2010), focusing in particular on controversies around user data access, ownership, and control. We characterise different actors in the Twitter data ecosystem: private and institutional end users of Twitter, commercial data resellers such as Gnip and DataSift, data scientists, and finally Twitter, Inc. itself; and describe their conflicting interests. We furthermore study Twitter's Terms of Service and application programming interface (API) as material instantiations of regulatory instruments used by the platform provider and argue for a more promotion of data rights and literacy to strengthen the position of end users.

Keywords

Social media, Twitter, big data, users, platforms, regulation

Contents

1. The Big Data Moment

> *[D]ata is not free, and there's always someone out there that wants to buy it. As an end-user, educate yourself with how the content you create using someone else's service could ultimately be used by the service-provider. (Jud Valeski, CEO of Gnip, quoted in Steele, 2011, para 19)*

> *There are significant questions of truth, control, and power in Big Data studies: researchers have the tools and the access, while social media users as a whole do not. Their data were created in highly context-sensitive spaces, and it is entirely possible that some users would not give permission for their data to be used elsewhere. (boyd & Crawford, 2012, p. 12)*

Talk of Big Data seems to be everywhere. Indeed, the apparently value-free concept of 'data' has seen a spectacular broadening of popular interest, shifting from the dry terminology of lab coat-clad scientists to the buzzword *du jour* of marketers. In the business world, data is increasingly framed as an economic asset of critical importance, a commodity en par with scarce natural resources (Backaitis, 2012; Rotella, 2012), while in context with "open" public sector data there is a growing debate about digital information as an enabler of growth, transparency, and civic engagement.

It is social media that has most visibly brought the Big Data moment to media and communication studies, and beyond it, to the social sciences and humanities. Social media data is one of the most important areas of the rapidly growing data market (Manovich, 2012; Steele, 2011). Massive valuations are attached to companies that directly collect and profit from social media data, such as Facebook and Twitter, as well as to resellers and analytics companies like Gnip and DataSift. The expectation attached to the business models of these companies is that their privileged access to data and the resulting valuable insights into the minds of consumers and voters will make them irreplaceable in the future. Analysts and consultants argue that advanced statistical techniques will allow the detection of on-going communicative events (natural disasters, political uprisings) and the reliable prediction of future ones (electoral choices, consumption).

These predictions are made possible through cheap networked access to cloud-based storage space and processing power, paired with advanced computational techniques to investigate complex phenomena such as language sentiment (Thelwall, Buckley, & Paltoglou, 2011; Thelwall, to appear), communication during natural disasters (Sakai, Okazaki, & Matsuo, 2010), and information diffusion in large networks (Bakshy, Rosenn, Marlow, & Adamic 2012). Such methods are hailed as superior tools for the accurate modelling of social processes and have a growing base of followers among the proponents of "digital methods" (Rogers, 2009) and "computational social science" (Lazer et al., 2009). While companies, governments, and other stakeholders previously had to rely on vague forecasts, the promise of these new approaches is ultimately to curb human unpredictability through information. The traces created by the users of social media platforms are harvested, bought, and sold; as an entire commercial ecosystem is forming around social data, with analytics companies and services at the helm (Burgess & Bruns, 2012; Gaffney & Puschmann, to appear).

Yet, while the data in social media platforms is sought after by companies, governments and scientists, the users who produce it have the least degree of control over "their" data. Platform providers and users are in a constant state of negotiation regarding access to and control over information. Both on Twitter and on other platforms, this negotiation is conducted with contractual and technical instruments by the provider and with ad-hoc activism by some users.

The complex relationships among platform providers, end users, and a variety of third parties (e.g., marketers, governments, researchers) further complicates the picture. These nascent conflicts are likely to deepen in the coming years, as the value of data increases while privacy concerns mount and those without access feel increasingly marginalised.

Our paper approaches Twitter through the lens of "platform politics" (Gillespie, 2010), focusing in particular on controversies around user data access, ownership, and control. We characterise different actors in the Twitter ecosystem: private and institutional end users of Twitter, commercial data resellers such as Gnip and DataSift, data scientists, and finally Twitter, Inc. itself; and describe their conflicting interests. We furthermore study Twitter's Terms of Service and application programming interface (API) as material instantiations of regulatory instruments used by the platform provider and argue for a more promotion of data rights and literacy to strengthen the position of end users.

2. Twitter and the Politics of Platforms

The creation of social media data is governed by an intricate set of dynamically shifting and often competing rules and norms. As business models change, the emphasis on different affordances of the platform changes, as do the characteristics of the assumed end user under the aspects of value-creation for the company. Twitter has been subject to such shifts throughout its brief history, as the service adapts to a growing user community with a dynamic set of needs.

In this context, there has been a recent critique of a perceived shift from an 'open' Internet (where open denotes a lack of centralised control and a divergent, rather than convergent, software ecosystem), toward a more 'closed' model with fewer, more powerful corporate players (Zittrain, 2008). Common targets of this critique include Google, Facebook, and Apple, who are accused of monopolising specific services and placing controls on third-party developers who wish to exploit the platforms or contribute applications which are not in accordance with the strategic aims of the platform providers. In Twitter's case, the end of the Web 2.0 era, supposedly transferring power to the user (O'Reilly, 2005), is marked by the company's shift to a more media-centric business model relying firstly on advertising and corporate partnerships and, crucially for this paper, on reselling the data produced collectively by the platform's millions of users (Burgess & Bruns, 2012; van Dijck, 2012). This shift has been realised materially in the architecture of the platform—including not only its user interface, but also the affordances of its API and associated policies, affecting the ability of third-party developers, users, and researchers to exploit or innovate upon the platform.

There have been several recent controversies specifically around Twitter data access and control:

- the increasing contractual limitations placed on content through instruments such as the Developer Display Requirements (Twitter, 2012c), that govern how tweets can be presented in third-party utilities, or the Developer Rules of the Road (Twitter, 2012b), that forbid sharing large volumes of data;
- the requirement for new services built on Twitter to provide benefits beyond the service's core functionality;
- actions against platforms which are perceived by Twitter to be in violation of these rules, e.g. Twitter archiving services such as 140Kit and Twapperkeeper.com, business analytics services such as PeopleBrowsr, and aggregators like IFTTT.com;
- the introduction of the Streaming API as the primary gateway to Twitter data, and increasing limitation placed on the REST API as a reaction to growing volumes of data generated by the service;
- the content licensing arrangements made between Twitter and commercial data providers Gnip and Datasift (charging significant rates for access to tweets and other social media content); and
- the increasing media integration of the service, emphasizing the role of Twitter as "an information utility" (Twitter co-founder Jack Dorsey, quoted in Arthur, 2012).

In the following, we relate these aspects to different actors with a stake in the Twitter ecosystem.

3. Conflicting Interests in the Twitter Ecosystem

Lessig (1999) names four factors shaping digital sociotechnical systems: the market, the law, social norms, and architecture (code and data). The regulation of data handling by the service provider through the Terms of Service and the API is of particular interest in this context. As outlined above, Twitter seeks to regulate use of data by third parties through the Terms and the API, assigning secondary roles to the law (which the Terms frequently seek to extend) and social norms (which are inscribed and institutionalised in various ways through both the interface and widespread usage conventions).

3.1 Twitter, Inc.

Platform providers like Twitter, Inc. have a vested interest in the information that flows through their service, and as outlined above, these interests have become more pronounced over time, as the need for a plausible business model has grown more urgent. The users' investment of time and energy is the foundation of the platform's value and therefore growing and improving the service is of vital importance. In the case of Twitter, this strategy is exemplified by the changes made to the main page over the years. Whereas initially Twitter asked playfully, "What are you doing?," this invitation has long since been replaced by a more utilitarian and consumer-oriented exhortation to "Find out what's happening, right now, with the people and organizations you care about," stressing Twitter's relevance as a real-time information hub for business and the mainstream media.

Twitter's business strategy clearly hinges strongly on establishing itself as an irreplaceable real-time information source and on playing a vital part in the corporate media ecosystem of news propagation. Under its current CEO, Dick Costolo, Twitter has moved firmly towards an ad-supported model of "promoted tweets" similar to Google's AdWord model. Exercising tighter control over how users experience and interact with the service than in the service's fledgling days is a vital component of this strategy.

Data is a central interest of Twitter in its role as a platform provider, not solely because it aims to monetise information directly, but because the value of the data determines the value of the company to potential advertisers.

Increasing the relevance of Twitter as a news source is crucial, while maintaining a degree of control over the data market that is evolving under the auspices of the company.

3.2 End-users

Twitter's end users are private citizens, celebrities, journalists, businesses, and organisations; in other words, they can be both individuals and collectives, with aims that are strategic, casual, or a dynamic combination of both. What unites these different stakeholders is that they have an interest in being able to use Twitter free of charge and that data is merely a by-product of their activity, but not their reason for using the platform. They do, however, have an interest in controlling their privacy and being able to do the same things with their information that both Twitter and third-party services are able to do. While the Terms spell out certain rights that users have and constraints that they are under, the rights can only be exercised through the API, while the constraints are enforced by legal means (Beurskens, to appear).

End users have diverse reasons for wanting to control their data, including privacy concerns, impression management, fear of repressive governments, the desire to switch from one social media service to another, and curiosity about one's own usage patterns and behaviour. Giving users the ability to exercise these rights not only benefits users, but also platform providers, because it fosters trust in the service. The perception that platform providers are acting against users' interests behind their back can be successfully countered by implementing tools that allow end users greater control of "their" information.

3.3 Data traders and analysts

Both companies re-selling data under license from Twitter and their clients have interests which are markedly different from those of the company and platform end users. While Twitter seeks long-term profits guaranteed by controlled access to the platform and growing relevance, and end users may want to guard their privacy and control their information while being able to use a free service, data traders want access to vast quantities of data that allow them to model and predict user behaviour on an unprecedented scale. Access to unfiltered, real-time information (provided to them in the form of the

Streaming API) is vital, while to their clients the predictive power of the analytics is important. Neither is very concerned with the interests of end users, who are treated similarly to subjects in an experiment of gigantic proportions. Privacy concerns are backgrounded as they would reduce the quality of the analytics, and they are effectively traded for free access to the platform. What is also neglected is the ability to access historical Twitter data, as businesses by and large want to monitor their current performance, with only limited need to peer into the past.

A key aim of data traders is to commodify data and to guard it carefully against infringers operating outside the data market. In an interview, data wholesaler Gnip's CEO Jud Valeskii returns the responsibility back on end users, recommending they educate themselves about the public and commodified status of the data generated by their personal media use:

> *Read the terms of service for social media services you're using before you complain about privacy policies or how and where your data is being used. Unless you are on a private network, your data is treated as public for all to use, see, sell, or buy. Don't kid yourself. (Valeski, quoted in Steele, 2011, para 27)*

Two things stand out in this statement: the claim that data on Twitter is public and the inference that because it is public, it should be treated as "for all to use, see, sell, or buy." The public-private dichotomy applies to Twitter data only in the sense that what is posted there is accessible to anyone accessing the Twitter website or using a third-party client (with the exception of direct messages and protected accounts). But the question of access is legally unrelated to the issue of ownership—rights to data cannot be inferred from technical availability alone, otherwise online content piracy would be legal. In the same interview, Valeski also consistently refers to platform providers such as Twitter as "publishers" and warns of "black data markets."

4. Terms of Service and API as Instruments of Regulation

Since its launch in March 2006, Twitter has steadily added documents that regulate how users can interact with its service. In addition to the Terms

(Twitter, 2012a), two items stand out: the Developer Rules of the Road (Twitter, 2012b) and the Developer Display Requirements (Twitter, 2012c), which were added to the canon in September 2012. Twitter's Terms have changed considerably since Version 1, published when the platform was still in its infancy. In relation to data access, they lay out how users can access information, what rights Twitter reserves to the data that users generate, and what restrictions apply. Initially the Terms spell out the users' rights with respect to their data, i.e., each user's own personal content on the platform:

> *By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). (Twitter 2012a, para 5-1)*

This permission to use the data is supplemented with the permission to pass it on to sanctioned partners of Twitter:

> *You agree that this license includes the right for Twitter to make such Content available to other companies, organizations or individuals who partner with Twitter for the syndication, broadcast, distribution or publication of such Content on other media and services, subject to our terms and conditions for such Content use. (ibid, para 5-2)*

Third parties are also addressed in the Terms and encouraged to access and use data from Twitter:

> *We encourage and permit broad re-use of Content. The Twitter API exists to enable this. (ibid, para 8-2)*

However, the exact meaning of *re-use* in this context remains unclear, and reading the other above-mentioned documents, the impression is that data analysis is not the kind of re-use intended by the Terms. Neither is it made explicit whether the content referred to is still the users' own content or all data on the platform (i.e., the data of other users). Furthermore, it seems that it is no longer Twitter's users who are addressed, but third parties, as no referent is given. Reference to the API also suggests that a technologically savvy audience is addressed, rather than any typical user of Twitter.

The claim of encouraging broad re-use is further modified by the Developer Rules of the Road, the second document governing how Twitter handles data:

> *You will not attempt or encourage others to: sell, rent, lease, sublicense, redistribute, or syndicate access to the Twitter API or Twitter Content to any third party without prior written approval from Twitter. If you provide an API that returns Twitter data, you may only return IDs (including tweet IDs and user IDs). You may export or extract non-programmatic, GUI-driven Twitter Content as a PDF or spreadsheet by using 'save as' or similar functionality. Exporting Twitter Content to a datastore as a service or other cloud based service, however, is not permitted. (Twitter 2012b, para 8)*

Here, too, developers, rather then end-users are the implicit audience. Not only is the expression "non-programmatic, GUI-driven Twitter Content" fairly vague, the restrictions with regards to means of exporting and saving the data make the "broad re-use" that Twitter encourages in the Terms difficult to achieve in practice. They also stand in contradiction to the Terms which state:

> *Except as permitted through the Services (or these Terms), you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Content or Services. (Twitter 2012a, para 8-2)*

Thus, only by using the API and obtaining written consent from Twitter is it possible to redistribute information to others. This raises two barriers—requiring permission and having the technical capabilities needed to interact with the data—that must both be overcome, narrowing the range of actors able to do so to a small elite. In relation to this form of exclusion, boyd and Crawford (2012) speak of data "haves" and "have-nots," noting that only large institutions with the necessary computational resources will be able to compete. Studies such as those by Kwak, Lee, Park, and Moon (2010) and Romero, Meeder, and Kleinberg (2011) are only possible through large-scale institutional or corporate involvement, as both technical and contractual challenges must be met. While vast quantities of data are theoretically available via Twitter, the process of obtaining it is in practice complicated, and requires a sophisticated infrastructure to capture information at scale.

Actions such as the one against PeopleBrowsr, an analytics company that was temporarily cut off from access to the API, support the impression that Twitter is exercising increasingly tight control over the data it delivers through its infrastructure (PeopleBrowsr, 2012). PeopleBrowsr partnered with Twitter for over four years, paying for privileged access to large volumes of data, but as a result of its exclusive partnerships with specific data resellers, Twitter unilaterally terminated the agreement, citing PeopleBrowsr's services as incompatible with its new business model.

5. Data Rights and Data Literacy

Contemporary discussions of end user data rights have focused mainly on technology's disruptive influence on established copyright regimes, and industry's attempts to counter this disruption. Vocal participants in the digital rights movement are primarily concerned with copyright enforcement and Digital Rights Management (DRM), which, so the argument goes, hinder democratic cultural participation by preventing the free use, embellishment, and re-use of cultural resources (Postigo, 2012a, 2012b). The lack of control that most users can exercise over data they have themselves created in platforms such as Twitter seems a in some respects a much more pronounced issue.

Gnip's CEO Jud Valeski frames the "owners" of social media data to be the platform providers, rather than end users, a significant conceptual step forward from Twitter's own characterization, which endows the platform with the licence to reuse information, but frames end users as its owners (in Steele, 2011). Valeski's logic is based on the need to legitimise the data trade—only if data is a commodity, and if it is owned by the platform provider rather than the individual users producing the content, can it be traded. It furthermore privileges the party controlling the platform technology as morally entitled to ownership of the data flowing through it.

Driscoll (2012) notes the ethical uncertainties surrounding the issues of data ownership, access, and control, and points to the promotion of literacy as the only plausible solution:

> *Resolving the conflict between users and institutions like Twitter is difficult because the ethical stakes remain unclear. Is Twitter ethically bound to explain its internal algorithms and data structures in a language that its users can understand? Conversely, are users ethically bound to learn to speak the language of algorithms and data structures already at work within Twitter? Although social network sites seem unlikely to reveal the details of their internal mechanics, recent 'code literacy' projects indicate that some otherwise non-technical users are pursuing the core competencies necessary to critically engage with systems like Twitter at the level of algorithm and database. (p. 4)*

In the current state, the ability of individual users to effectively interact with "their" Twitter data hinges on their ability to use the API, and on their understanding of its technical constraints. Beyond the technical know-how that is required to interact with the API, issues of scale arise: the Streaming API's approach to broadcasting data as it is posted to Twitter requires a very robust infrastructure as an endpoint for capturing information (see Gaffney & Puschmann, to appear). It follows that only corporate actors and regulators—who possess both the intellectual and financial resources to succeed in this race—can afford to participate, and that the emerging data market will be shaped according to their interests. 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.

6. References

Arthur, C. (2012). Twitter too busy growing to worry about Google+, says Dorsey. *Guardian.co.uk*. Retrieved from http://www.guardian.co.uk/technology/2012/jan/23/twitter-dorsey

Backaitis, V. (2012). Data is the New Oil. *CMS Wire*.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The Role of Social Networks in Information Diffusion. *Proceedings of the 21st International*

Conference on the World Wide Web (WWW '12) (pp. 1–10). New York, New York, USA: ACM Press. doi:10.1145/2187836.2187907

Beurskens, M. (to appear). Legal questions of Twitter research. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Burgess, J. & Bruns, A. (2012). Twitter archives and the challenges of 'Big Social Data' for media and communication research. *M/C Journal, 15*(5). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/561

boyd, d. & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society* 15(5), 662-679.

Driscoll, K. (2012). From punched cards to "Big Data": A social history of database populism. *communication +1*, 1, Article 4. Retrieved from http://scholarworks.umass.edu/cpo/vol1/iss1/4

Gaffney, D., & Puschmann, C. (2012). Game or measurement ? Algorithmic transparency and the Klout score. *#influence12: Symposium & Workshop on Measuring Influence on Social Media* (pp. 1–2). Halifax, Nova Scotia, Canada.

Gaffney, D., Puschmann, C. (to appear). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347-364.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter , a social network or a news media? Categories and Subject Descriptors. *Proceedings of the 19th International Conference on the World Wide Web (WWW '10)* (pp. 591–600). Raleigh, NC.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Lessig, L. (1999). *Code and other laws of cyberspace*. New York, NY: Basic Books.

Manovich, L. (2012). Trending: The promises and the challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460-475). Minneapolis: University of Minnesota Press.

O'Reilly, T. (2005). What is Web 2.0? Design patterns and business models for the next generation of software. *O'Reilly Network*. Retrieved from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

PeopleBrowsr (2012). PeopleBrowsr wins temporary restraining order compelling Twitter to provide firehose access. Retrieved from http://blog.peoplebrowsr.com/2012/11/peoplebrowsr-wins-temporary-restraining-order-compelling-twitter-to-provide-firehose-access/

Postigo, H. (2012a). Cultural production and the digital rights movement. *Information, Communication and Society*, *15*(8), 1165-1185.

Postigo, H. (2012b) *The digital rights movement*. Cambridge, MA: MIT Press.

Rogers, R. A. (2009). *The end of the virtual: Digital methods*. Amsterdam, the Netherlands: Amsterdam University Press.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. *Proceedings of the 19th World Wide Web Conference* (pp. 695–704). New York, NY: ACM.

Rotella, P. (2012). Is Data The New Oil? *Forbes*. Retrieved October 5, 2012, from http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users. *Proceedings of the 19th International Conference on the World Wide Web (WWW '10)* (pp. 1–10). New York, NY: ACM Press. doi:10.1145/1772690.1772777

Steele, J. (2011). Data markets aren't coming, they're already here. *O'Reilly Radar*. Retrieved from http://radar.oreilly.com/2011/01/data-markets-resellers-gnip.html

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science*, *62*(2), 406–418. doi: 10.1002/asi.21462

Thelwall, M. (to appear). Sentiment Analysis and Time Series with Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt & C. Puschmann (eds.), *Twitter and Society*. New York, NY: Peter Lang.

Twitter (2012a). Terms of Service. Retrieved from http://twitter.com/tos.

Twitter (2012b) Rules of the Road. Retrieved from https://dev.twitter.com/terms/api-terms.

Twitter (2012c) Developer Display Requirements. Retrieved from https://dev.twitter.com/terms/display-requirements.

van Dijck, J. (2011). Tracing Twitter: The rise of a microblogging platform. *International Journal of Media and Cultural Politics*, *7*(3), 333-348.

Zittrain, J. (2008). *The future of the Internet and how to stop it*. New Haven, CT: Yale University Press.