Playing with digital byproduct data

An indicative example

David Beer

It is fairly common knowledge that all sorts of everyday interactions and engagements with media are captured as byproduct data.[1] It is probably also fairly well-known that such data is routinely harvested by capitalist organisations which then use it in an attempt to predict things about us.[2] However, the ways in which this byproduct data recursively feeds back into Web culture has received much less attention.[3]

One of the developments typical of what is often referred to as social media or Web 2.0 are the relatively common invitations and opportunities to *play with the data*. Often these forms of data-play are founded upon the creation of visualisations of various types. On some occasions these might be invitations to play with data generated by your own activities or those of our social networks – the opportunities to visualise and compare musical tastes on the last.fm 'playground' is a good example of this as are the many 'apps' that allow us to analyse friendship groups on Facebook.[4] Other applications ask that you bring your own data and provide the means for visualising the content in new ways; the creation of word clouds on the well-known wordle.com application would be a prominent example of this. Finally, there are a range of resources that provide both a data source and the means with which to analyse it. An illuminating example of this final category would be the Web resource wefeelfine.org. This site trawls the Web and captures content where the phrases 'I feel' or 'I am feeling' have been used. Wefeelfine.org provides real-time insights into emotional responses which can then be used to create various types of visualisations of the harvested data along the lines of gender, age, location and the like. We might raise questions about the nature of such data but what is clear is that there are various kinds of visual social research at play within these online resources.

In this review I have selected a web resource that falls into this final category: tweetolife.com. The chosen web resource provides both the data set and the means of analysis. Tweetolife is used here as an indicative and suggestive example of what is becoming possible. As such it is hoped that this review will stimulate some reflection upon how such digital-byproduct data and Web-based analytical techniques might fall into the critical gaze of media studies.[5]

If we look at the biographies of some of the applications mentioned in the opening paragraph we can quickly reveal the depth of different interests involved in creating these resources. In broad terms we might observe that computer scientists, designers, freelance visualisers, and commercial data solution companies are engaged in professional forms of visual studies. These professional forms of analysis

complement the growing sets of lay and amateur visualisation practices that are a common part of much contemporary media. The data visualisation expert Nathan Yau runs the popular data visualisation site flowingdata.com. In his recent book he suggests that visualisation techniques have moved outside of their origins within the academy and have become a fixture of popular culture.[6] As a result, he argues, the activities of statisticians and computer scientists have become the basis upon which some popular visualisation techniques and applications have moved into the cultural mainstream. Yau claims that 'somewhere in between journalism and art, visualization has also found its way into popular entertainment'.[7]

The chosen application Tweetolife carries the promising, if possibly a bit hopeful, tag line 'the science of human life in Twitter Messages'. According to the available background information the Tweetolife application (or 'demo') is:

> the result of a study that was carried out at the Language, Interaction and Computation Laboratory at the University of Trento in Italy. We looked at the daily patterns of life in Twitter messages (tweets), and we present the differences in the contents of tweets according to the gender of the users and time of the day.[8]

This application then is a product of an academic study and made available for users to play with the data set extracted from Twitter postings. Here the products of these academic practices – particularly in computer science and informatics – cross over into popular culture via the possibilities of media-based dissemination. According to the information available through the University of Trento web resources:

> [t]he Language, Interaction and Computation Laboratory (CLIC) is an interdisciplinary group of researchers interested in studying verbal and non-verbal communication. Research in the laboratory uses both computational and cognitive methods.[9]

Indeed a key reason for selecting Tweetolife from amongst the various similar applications is that it provides an example of the interface between academic statisticians, social scientists, and this broader interest in a type of lay visual studies that is typical of contemporary Web cultures.

The project that culminated in the Tweetolife application drew upon a fixed data set – that is to say that Tweetolife is not a real-time research tool, rather that it works from a data set extracted during a defined period of time (as such it does not operate in the same way as other real-time data harvesting technologies). The creators of Tweetolife describe the process as follows:

> [w]e analyzed millions of tweets collected by researchers from the University of Edinburgh between November 2009 and February 2010. For gender differences, we separated the tweets into two subsets as male and female tweets by using the first names of the Twitter users. For hourly differences, we grouped the tweets according to the time of the day they were posted (in each user's local time).[10]

The data was accumulated during this four-month period and the 'Twitter API' was used to capture the 'entire stream' of Twitter posts.[11] The accumulated data was then tagged with the local time and with the gender of the content creator. As the analysis below indicates the tagging of the data with only information about the time of posting and the gender of the person generating it is quite limiting in scope; we might also ask questions about the accuracy of such tagging processes.[12]

CLIC notes the lack of a public data resource for Twitter and claim to have the

> first such corpus collected over a period of two months using the Twitter streaming API. Our corpus contains 97 million tweets.[13]

Clearly this is a substantial set of data that records a vast quantity of ordinary interactions – which in turn speaks directly to calls in the digital humanities to archive and capture such data for historical use. Indeed, it is this vast data set that makes Tweetolife an example worth attention. The data accumulated during this short sample period is already quite staggering; we are instantly forced to reflect on how difficult it would be to accumulate a comparable resource through more traditional social science methods; it is also suggestive of the broader problems of data loss and the possibilities of archiving for future analysis and historical insights.[14]

We might of course question what it is that we can extract from such online interactions and how seriously we can take the content of the communication. However it is that we might feel about such content this is clearly a vast social data set that has the potential to inform and shape social and cultural research. This is the type of digital data set that social scientists have been expressing an increasing desire to obtain and use.[15] In this instance the data provided by Tweetolife can be analysed at what might be understood to be a quantitative or aggregate level but it is worth noting that elsewhere applications allow broad trends to be analysed at the level of particular postings. Thus these new forms of data analysis are often able to work across quantitative and qualitative forms of analysis.[16]

We have a sense of the size of the set of data used by CLIC and the type of analytical and technical processes that underpin Tweetolife. The question arises about what this application can allow us to do with this vast digital data set. The

various analytic options of Tweetolife are all set up to allow the user to enter their own chosen search term(s). The site itself is separated into two parts: the first deals with 'gender differences', the second with 'hourly differences'. Despite the flexibility in search terms our analysis is limited to these two possibilities. We will begin with gender differences.

The most basic feature of Tweetolife requires a single search term that is then used to generate a visualisation indicating the gender divisions in the content. Figure 1 was created by using the search term 'music'. The visualisation shows the gender breakdown of the content mentioning music. Immediately below the total percentage of content Figure 1 also shows the words most commonly associated with music by gender. Figure 1 appears to reveal that music is a fairly gender neutral topic in terms of the quantity of the posts; however, there would seem to be some notable differences revealed here in the way that men and women talk about music. We might be hesitant to draw any conclusions from this. Again, perhaps the usefulness of this particular application is in how suggestive it is of the other analytical possibilities rather than in the specifics of what it reveals.

Fig. 1: A visualisation created on Tweetolife using the search term 'music' to show the gender distribution of comments on Twitter containing this key term.

It is possible on Tweetolife to look further at these types of gender orientations by comparing the volume of content for a range of combined search terms. To give a sense of how this might be used seven search terms were entered to create Figure 2. Figure 2 compares the volume of gender-based content for the following search terms: music, film, TV, comedy, theatre, opera, and sport. Figure 2 does not reveal

the comparative volume of content for these different terms; rather, it shows the percentage of all the content that mentions each term and shows the gender split of that content. For example, Figure 2 shows that music, TV, theatre, and opera all have an even gender split whereas film, comedy, and particularly sport have a higher percentage of posts by male content creators.

Fig. 2: The percentage of content on various search terms broken down by gender.

We might be concerned that in most cases Figure 2 tends to be dominated by content created by men but further searches reveal that this appears to be a product of the types of search terms chosen.

It is possible to gain more of a comparative perspective on the volume of content from the time-based analysis that Tweetolife provides. In this instance this is not divided by gender but instead shows the relative volume of content posted across a 24-hour cycle (the content was tagged with the local time). This provides the user with insights into the everyday flows of conversations on topics and is suggestive of how these interactions fit into ordinary routines. The visualisations created are simple line graphs based upon the number of occurrences of that topic per million Twitter posts. What this provides then is not a simple indication of the volume of posts on a particular topic across a 24-hour cycle; rather, it shows the relative amount of posts concerned with that topic at different times of the day. Although the actual number of Twitter postings may be lower at 03:00h than at 15:00h the topic may show an increase in the line graph because that topic may have had a higher share of the postings at that particular time of day. As such the line graphs

generated show the importance or relevance of topics at different times of day within the Twitter network.

In Figures 3 and 4 I have again used the same search terms: music, film, TV, comedy, theatre, opera, and sport. Figure 3 shows the relative volume of postings on these various topics based upon the time at which they were created; it visualises the comparative volumes of posts on these various search topics; it reveals, for instance, that music is continuously the most popular topic and that there is comparatively very little content discussing comedy, theatre, or opera (as we might expect) – although we see that comedy does slightly gain in importance in the content generated late in the evening and in the very early hours of the morning. Figure 3 reveals the temporal patterns of importance of the content associated with these seven chosen cultural spheres. In most instances we see a notable increase in relative quantity of content in the early morning and in the evening. This is likely to be because the topics reflect pastimes that predictably appear to gain in importance around typical leisure times of the day. This is most pronounced in the case of TV, with a much higher number of occurrences per million Twitter postings on this topic around typical leisure times, particularly in the evening.

Fig. 3: The occurrences per million Twitter posts for selected topics during a 24-hour cycle.

We might wonder why we are seeing these variations in content across the daily cycle or more generally we may just wish to get a sense of the types of discussions that are occurring around these selected topics. In order to develop such insights it is possible to follow the content so as to reveal the words commonly associated

with each separate search term at different times of the day. Each hourly interval can be examined on each of the lines. To give an example Figure 4 is the same graph as Figure 3, but in this instance I have highlighted the point at 19:00h on the TV content line. Figure 4 shows the list of words commonly associated with TV in that time frame. This not only shows us the types of discussions that are occurring but can also be used to help with understanding the types of temporal variations in content that we have observed. The box generated for 19.00h indicates that the discussion of TV is commonly associated with the following words: relaxing, primetime, glue, Simpsons, and mute. The presence of 'relaxing' in the words most commonly associated with TV might be used to suggest the part that TV plays in routine daily cycles – the other commonly used words perhaps indicating shared interests in particular TV shows (or with muting the TV in order to communicate). By tracking across the peak in the discussion of TV that occurs in the evening it is possible to get a sense of the nature of this relative increase in content. If we track from 18:00h to 23:00h we find that the word 'relaxing' is amongst the most common words associated with TV through this time period. These might seem like fairly banal observations but again, they are suggestive of the way in which this byproduct data provides access into the everyday worlds of those within the network.

Fig. 4: As in Figure 3, but in this case the TV line has been selected at the time 19:00h to reveal the key terms associated with TV during that time period.

It is reasonable to be critical about the scope of Tweetolife given that the analytics are limited to gender and time. We might be able to use an unlimited set of search

terms on any chosen topic but we are always forced to work with gender and time as our analytic focus. We might also be concerned that the visualisations created by Tweetolife are not particularly imaginative from an aesthetic standpoint – indeed the other examples I listed in the opening paragraph all produce much more dramatic visualisations than Tweetolife. The significance of Tweetolife is not so much about what this one example can offer, although this is still revealing in its use of such a significant data set; rather, it is about how suggestive this application is of the broader possibilities for incorporating such byproduct data and analytical approaches into social and cultural research.

We can imagine that researchers across the social sciences and humanities will see possibilities for using such data to explore events historically, to analyse responses to issues, to observe cultural trends as they develop, to open up aspects of everyday life that tend to be difficult to see, and so on. Not in the least we can imagine that seeing patterns in such data will inevitably produce new insights that might then be pursued using more traditional techniques. Tweetolife is just one example of a much broader range of visual Web applications that enable new types of data to be harvested and analysed in various ways. Those interested in media studies may be interested in the analytic possibilities but they might be just as interested in critically responding to these new visual tools and the way that they 'enact' the social world.[17] These tools could become powerful actors that redefine the appreciation of things like gender, emotions, places, politics, and the like. For this reason it is important that such visual representations are treated to rigorous interrogation.

The example described in this review suggests the potential for the social sciences and humanities – and particularly media studies – to collaborate and engage with the emergent Web-based cultures of visualisation. I have outlined one example here; there are many other applications and visuals that might draw our critical and analytic attention. Similarly there are myriad individuals involved in this culture and industry of visualisation who would be ripe for innovative collaboration; these would include computer and information scientists (like those behind Tweetolife), designers and artists, freelance data visualisers, companies providing analytic solutions for data inundation – and this is not to mention the many lay and ordinary Web users who are engaging in visual analytics of various types as they mash data together and develop their own software apps. The driving force for a good proportion of this activity is play but this is a microcosm of a much bigger set of developments that we might want to tap into for ideas or respond to with much needed critical interrogations. Tweetolife might be quite basic in the analytical opportunities it offers but it provides insights into a massive data set and forces us to imagine what else might happen, or might already be happening, at the intersection between digital byproduct data and media studies.

Notes

1. Abbott 2000.
2. Thrift 2005.
3. Beer 2013.
4. A recent example of one of these apps uses algorithmic analysis of postings and status updates to enable the user to observe the real-time emotions of their friendship group.
5. Beer & Burrows 2010, pp. 233-252.
6. Yau 2011.
7. Ibid.
8. The background information is available at http://www.tweetolife.com/about/about.html.
9. Access the information about CLIC at http://clic.cimec.unitn.it/.
10. http://www.tweetolife.com/about/about.html.
11. For an account see Petrović et al. 2011.
12. We might note that the use of name to allocate gender could be something of a problem.
13. Petrović et al 2010, p. 1.
14. For more on the politics of memory in new media see Dodge & Kitchin 2007.
15. Abbott 2000; Savage & Burrows 2007; Adkins & Lury 2009.
16. For a description of an application that allows digital data to be analysed across quantitative and qualitative dimensions see Beer 2012.
17. See Law 2004.

References

Abbott, A. 'Reflections on the future of sociology', *Contemporary Sociology*, Vol. 29, No. 2, 2000: 296-300.

Adkins, L. and Lury, C. 'What is the empirical?', *European Journal of Social Theory*, Vol. 12, No. 1, 2009: 5-20.

Beer, D. 'Using social media data aggregators to do social research', *Sociological Research Online*, Vol. 17, No. 3, 2012. http://www.socresonline.org.uk/17/3/10.html.

_____. *Popular culture and new media: The politics of circulation*. Basingstoke: Palgrave Macmillan, 2013.

Beer, D. and Burrows, R. 'Sociological imagination as popular culture' in *New social connections: Sociology's subjects and objects* edited by J. Burnett, S. Jeffers, and G. Thomas. Basingstoke: Palgrave Macmillan, 2010.

Dodge, M. and Kitchin, R. 'Outlines of a world coming into existence: pervasive computing and the ethics of forgetting', *Environment and Planning B: Planning and Design*, Vol. 34, No. 3, 2007: 431-445.

Law, J. *After method*. Cambridge: Routledge, 2004.

Petrović, S., Osborne, M. and Lavrenko, V. 'The Edinburgh Twitter Corpus', http://homepages.inf.ed.ac.uk/miles/papers/socmed10.pdf, 2010 (accessed 20 October 2011).

Savage, M. and Burrows, R. 'The Coming Crisis of Empirical Sociology', *Sociology*, Vol. 41, No. 6, 2007: 885-899.

Thrift, N. *Knowing capitalism*. London: Sage, 2005.

Yau, N. *Visualize this*. Oxford: Wiley, 2011.

About the author

David Beer, *University of York*