Digital Data Analysis, Public Engagement and the Social Life of Methods

Interim Report

Helen Kennedy, Giles Moss, Chris Birchall, Stylianos Moshonas

Introduction

People's web and social media use generates a vast new source of data which, it is suggested, can be analysed for new insights into what people think and feel, how they behave, and about the nature of social networks and relationships. *Digital Data Analysis, Public Engagement and the Social Life of Methods* aims to interrogate such bold epistemological claims about what the analysis of digital, social media or 'big' data might tell us. It does so by investigating the use of digital data analysis in practice. Working with three public sector organisations (two city councils and one local museum), our research is experimenting with different forms of digital data analysis in order to examine how they might help public organisations to know their publics better. Whilst we have some technical expertise and have consulted with an expert in social media insights at various points during the project, none of us is an expert in social media data analysis and insights and as such we may well resemble teams in public sector organisations attempting to grapple with these methods for the first time As we describe below, while digital data analysis may be valuable for public sector organisations in theory, our research suggests that the data, code and forms of expertise that it requires may be less accessible, open and well distributed in practice than is commonly supposed.

Progress with empirical research

Over the past three months, we have experimented with various tools and forms of digital data analysis. In terms of tools, we have used:

- the social data platform *DataSift* in order to gather data;
- *NodeXL, Gephi* and *Issuecrawler* to conduct network analysis in order to identify key 'nodes' in networks (key influencers, gateway nodes or groups) with whom our partners might want to engage, and to produce visualisations of these networks;
- *Overview* in order to analyse the content of social media interactions and identify what issues relevant publics are discussing.

These tools were selected because they fitted with the aims of our partner organisations and because they are free or low cost (a small fee is required in order to access data through *DataSift*) and so would be affordable for the organisations after the project's completion.

We asked our partners to identify two topics each on which we could focus our investigations. They came up with:

1. A photography exhibition about nature (Partner 1 (museum));
2. A set of online learning resources about art/museum education (Partner 1 (museum));
3. Council budget cuts (Partner 2 (council));
4. Health & well-being initiatives (Partner 2 (council));
5. The Tour de France in Yorkshire (Partner 3 (council));

6. The opening of a new city centre market (Partner 3 (council)).

Our partners identified approximately 20 keywords for each topic, to inform our searches. On the whole, searching with these terms has not resulted in us finding large quantities of data. However, the terms served as starting points for identifying other, more widely used terms. For example, we utilised the Flickr search API to find search terms that its community have used to tag images that are also tagged with the keywords supplied by Partner 1. This allows a picture to be built up of related keywords and keyword usage by people on Flickr (who we assume to be interested in images, the focus of our first topic). Searches were carried out for many of the search terms supplied and the result of these searches was a network of connected keywords which was then analysed in Gephi.

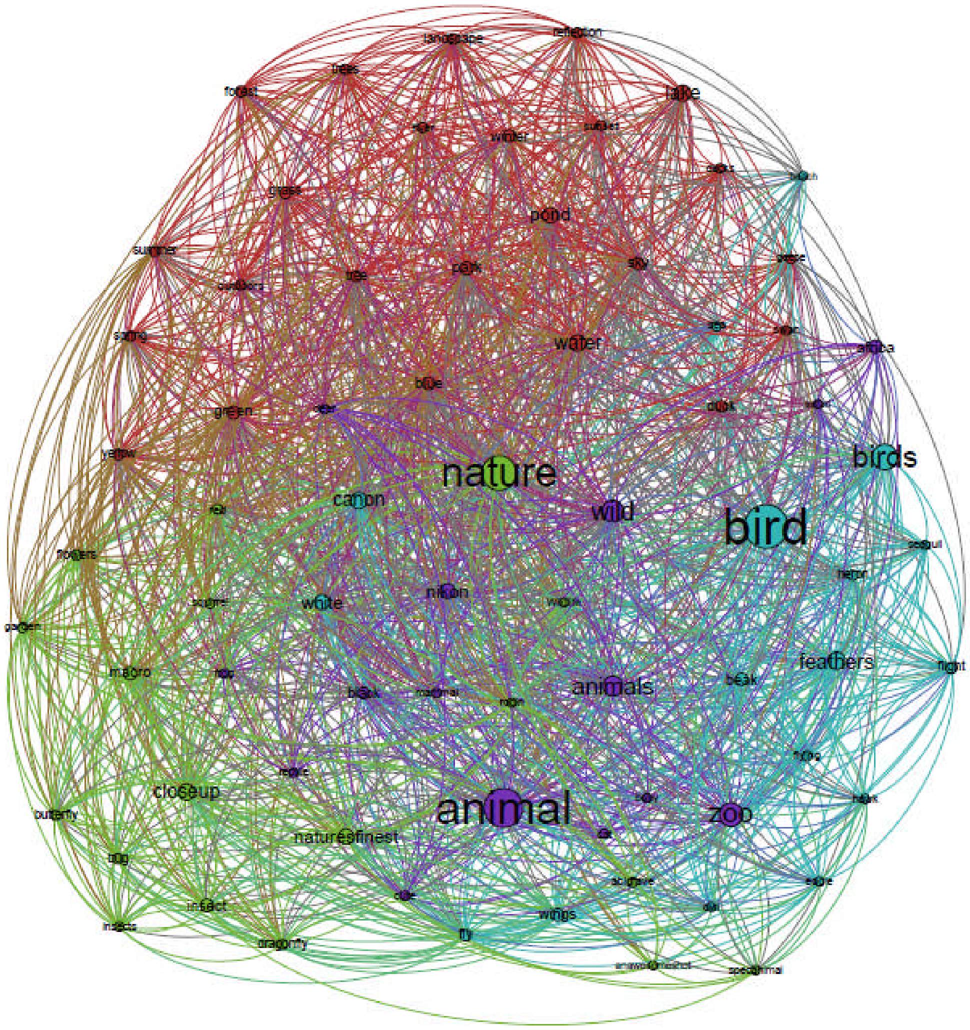

Figure 1: Network of connected Flickr keywords, seeded by the keyword 'animal'.

Analysing for 'degree centrality' (the number of connections one node has to other nodes), 'closeness centrality' (how close a node is to all other nodes – the sum of the shortest paths between nodes), 'betweenness centrality' (how often a node appears in the shortest path between other nodes and how important it is as a gateway) and for sub-communities within the network gave insights into which keywords were used to describe particular types of image content, as well as keywords that transcended these communities. For instance, it was observed that amongst images tagged with the word 'animal' a more important keyword, which transcended the sub communities of animal-related images, was 'nature'. Both 'animal' and 'nature' are keywords supplied by the partner, but our findings suggest that after including 'nature' in a search, there is no need to include ‘animals’, which in turn facilitates a more tightly targeted search for content related to natural photography. In this way, keywords could be chosen that allowed searches to be targeted more specifically at particular communities or broadened out to include a wider range of communities, depending upon the goals of the search.

Being local organisations, most partners are interested in geographically specific searches. A time zone filter can be applied to Twitter searches, as users have to specify a time zone when they sign up, but it is estimated by the expert social media insights consultant who is working with us on the project that approximately 90% of available data can be left out of a search if we seek only to gather data from sources known to be in a specific geographical area, because most users do not supply this data. For example, experiments were carried out to limit the data harvested by DataSift queries using geographical data, such as limiting the search to a 50-mile radius of a specified location or within a polygon representing a specified catchment area. However, these searches were unsuccessful; not because of a lack of talk in the area but because of a lack of geographical data attached to most contributions.

Our partners provided us with lists of their own social media accounts and key influencers already known to them. From this, we have generated a mapped network of websites for Partner 1, harvested using IssueCrawler and visualised using Gephi (see figure 2 below). This map was created by seeding the IssueCrawler search with URLs discovered during Google searches of the known influential names, and URLs and Twitter accounts supplied by Partner 1. Although the IssueCrawler data does not represent a social network, it does represent a set of connected nodes, some linked more strongly than others and some positioned more centrally in the network. Rather than looking for key influencers, the process is looking for ‘gateway nodes’ or nodes that are comprehensive or authoritative sources of information. The links and content are not specific to Partner 1’s identified topics, but this was generated from URLs that are relevant to the partner, so gives a picture of the interconnectedness of a network of relevant sites. This approach revealed the presence of a number of interesting new entities, including transport websites, with which staff at the partner museum could engage. This experiment suggests that broader searches which are generally partner-related may be more productive than narrow, topic-specific searches. We will develop similar visualisations for our other partners in order to explore this hypothesis further.

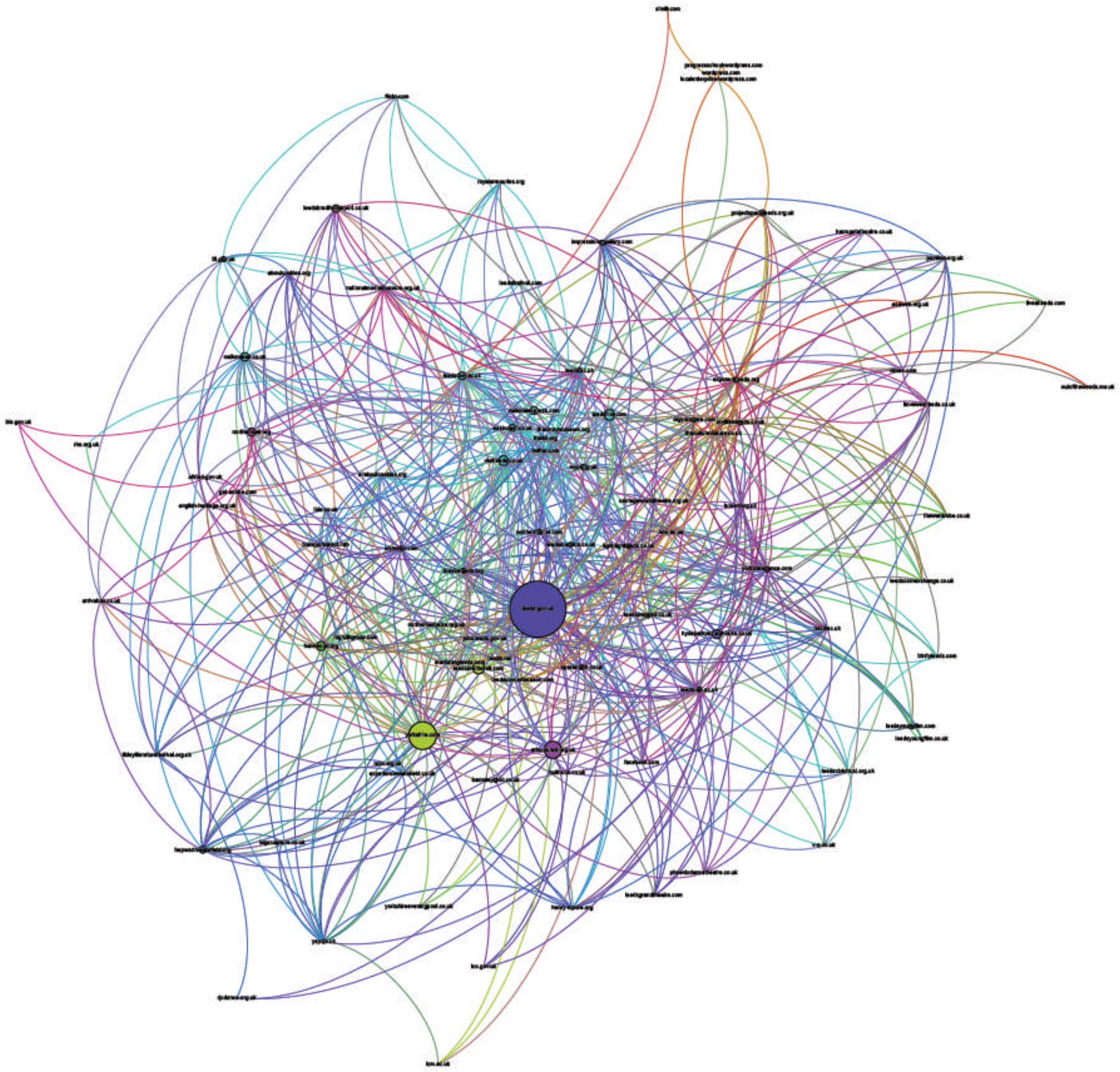

Figure 2: Visualisation of map of networked websites for Partner 1

Experiments such as those outlined in the previous paragraph were not in our original research proposal. However, difficulties in finding relevant data have meant that we have had to diversify our approaches and be creative with what we do. The use of free tools has been difficult, labour intensive and reliant on technical know-how. It is possible that the difficulties we have experienced accessing data could be a result of how DataSift works. We do not know how this black-boxed technology operates, what it harvests, where it searches and how far it reaches (although some indication of this is given within the platform), but it is possible that it circumscribes extensively what can be found and what cannot. However, it is also possible that we have not been able to find social media conversations about our partners' identified topics because not much is being said about these topics on social media platforms. It may be that people are simply not talking about council issues, for example, on Twitter and Facebook. We have identified some such conversations through manual searches on city-based forums and in the comments sections of local news websites, which data gathering tools like DataSift are just not extensive enough to reach. It may be, then, that other methods, such as manual searches, Google alerts, or focus groups, are better for understanding and engaging with publics in these specific cases.

Because of these limitations in finding data, we have made limited progress with conversation analysis in Overview; we hope to do more of this in the next three months. And also because of these limitations and an interest from one of our partners, we will trial two commercial social media data analysis platforms, Meltwater Buzz and Brandwatch, in June. We will reflect on the resourcing pros and cons of these paid-for services compared to free-but-complex tools like NodeXL and Gephi, and also compare results, in order to examine whether topics such as those chosen by our partners are discussed on social media platforms. We will also reflect on the level of expertise needed to use these tools and to make sense of the data that they produce, and on whether required levels of expertise may prohibit the kinds of socially inclusive analytics, or analytics for the public good, that this project seeks to facilitate.

As can be seen above, our research is most advanced with Partner 1 at this stage. We have generated visualisations of networks for them, which identify new possible 'influencers' and communities with whom to engage. They are also the most keen to learn how to use the tools with us, so that they can continue to use them after our project is complete. They have installed the relevant software in their workplace and have experimented with using DataSift to find data. We have had frequent meetings with them and provided them with ample guidance on how to use the tool. Engagement in our project may be most straightforward for them as they represent a comparatively small organisation and are not constrained by the need for extensive consultation. We have also held a number of meetings with Partner 2, and we hope that at our next meeting with them we will be able to provide them with some data. Partner 3 joined the project later; we have met infrequently, but the process with them is speeded up through our learning with other partners. We will meet in June and hope to provide them with data. The size and complexity of these two partner organisations makes it difficult for them to use the tools alongside us. Therefore we have proposed a one-day workshop at the University of Leeds towards the end of the project, during which partner representatives (up to five from each organisation) will experiment with the tools. We will provide them with a report on our experience of tool use at that meeting.

Reflections / preliminary thoughts on research questions

1. *How is data generated from social media and web use currently used by public-sector organisations to measure or enhance public engagement?*

This question was explored during a scoping study prior to this pilot project, the results of which can be found on the CCN+ website (http://www.communitiesandculture.org/files/2013/04/Scoping-report-Leeds-Suggestions.pdf).

2. *How might new forms of digital data analysis help these organisations to enhance their understanding and engagement of their publics?*

Our research suggests that the data, tools and forms of expertise required in order to conduct digital data analysis are less open and accessible in practice than is commonly supposed. Despite the rhetoric around digital data analysis, we have found that:

(a) data is more difficult to access than often assumed; in commercial social media monitoring companies, very broad data sets are accessed to generate 'findings', not the kinds of specific data sets which perhaps we and our partners imagined we would find;

(b) some conversations may not be happening on the main social media platforms that available tools search, so it may be that automated social media data analysis is not the best method to access such conversations;

(c) digital data analysis with freely available tools is labour intensive;

(d) digital data analysis with freely available tools is reliant on technical expertise;

(e) expertise is also required in order to make sense of – and act on – data that is produced. (This may be as much the case for commercial tools as for free ones.)

3. *In what ways and to what extent are those methods known as social media monitoring, web analytics and big data socially constructed?*

Digital data analysis is never just 'technical', but is dependent upon various assumptions and conjectures about the object of analysis and methodological decisions about how to capture and analyse it. Our processes of choosing keywords and of refining keywords serve to 'shape' what data will be found and what will not. But such assumptions and decisions are not all made by the analyst; they are also embedded in the code and algorithms of digital data applications and services. More needs to be known about how these function. Digital data analysis may most accurately be described as a diverse association or assemblage (Latour 2005), which combines various social and technical elements: data, platforms, computers, code, algorithms, procedures, values, assumptions, conjectures and so on.

4. *How can actors who do not have the economic means to pay for digital data and who want to use it for the public good access it?*

We raised this question in our report from the summer scoping study. In our view, it is important that digital data analysis remains open, both in terms of access to data and technologies and the transparency of its processes. In democratic terms, it is important that digital data analysis remains open to varied public purposes and uses, not just those that we might expect it to be adopted for (that is, for marketing and advertising in the commercial sector or for policing and national security). However, given question marks about the accessibility of the data and the tools, digital data analysis may be less well distributed than is often supposed. I It may be that paid-for, commercial services are easier to use than free tools, but these services can be expensive, resulting in a tradeoff between usability and affordability. Commercial services may also be less open about how their software and algorithms operate and about the types of methodological decisions and assumptions they make

Latour, B. (2005) *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.