UNIVERSITY OF LEEDS

Digital Data Analysis:

Guide to tools for social media & web analytics and insights

July 2013

From the research project *Digital Data Analysis, Public Engagement and the Social Life of Methods*

Helen Kennedy, Giles Moss, Chris Birchall, Stylianos Moshonas, Institute of Communications Studies, University of Leeds

Funded by the EPSRC/Digital Economy Communities & Culture Network + (http://www.communitiesandculture.org/)

Contents

Digital Data Analysis: 1
1. Free online tools 2
1.1. Social Mention (http://www.socialmention.com/) 2
1.2. Topsy (http://topsy.com/) 3
1.3. Klout (http://klout.com/home), Kred (http://kred.com/), Peer Index (http://www.peerindex.com/) 3
1.4. TweetReach (http://tweetreach.com/) 4
1.5. SentiStrength (http://sentistrength.wlv.ac.uk/) 5
1.6. Other online tools 5
2. Digital Data Analysis project tools 6
2.1. DataSift (http://datasift.com) 6
2.2. IssueCrawler (www.issuecrawler.net) 14
2.3. NodeXL (http://nodexl.codeplex.com/) 17
2.4. Gephi (https://gephi.org/) 23
2.5. Overview (https://www.overviewproject.org/) 29
3. Commercial platforms & services 33
3.1. Meltwater Buzz (http://www.meltwater.com/products/meltwater-buzz-social-media-marketing-software/) 33
3.2. Brandwatch (http://www.brandwatch.com/) 38
3.3. Other commercial platforms & services 44

Introduction

Digital Data Analysis was a short, six month, exploratory research project carried out by staff at the Institute of Communications Studies, the University of Leeds (ics.leeds.ac.uk/), and funded by the Digital Economy Communities and Culture Network+ (http://www.communitiesandculture.org/).

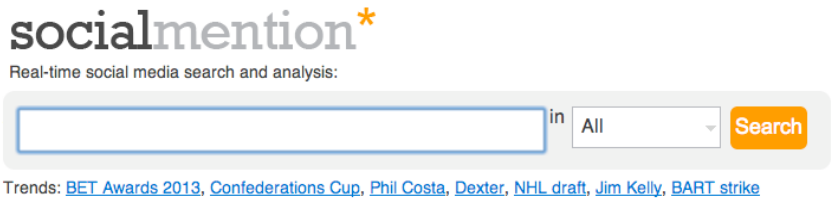
The project experimented with different tools for digital/social media data analysis in order to examine how the data they produce might help public organisations to know and engage their publics better. This document provides brief summaries of the tools with which we experimented during the course of the project. It emphasises the usefulness or otherwise of the tools to public sector organisations for the purposes of public engagement. As such, it does not represent a comprehensive review of the full capabilities of each tool. It also highlights some other tools, many of which are available freely online.

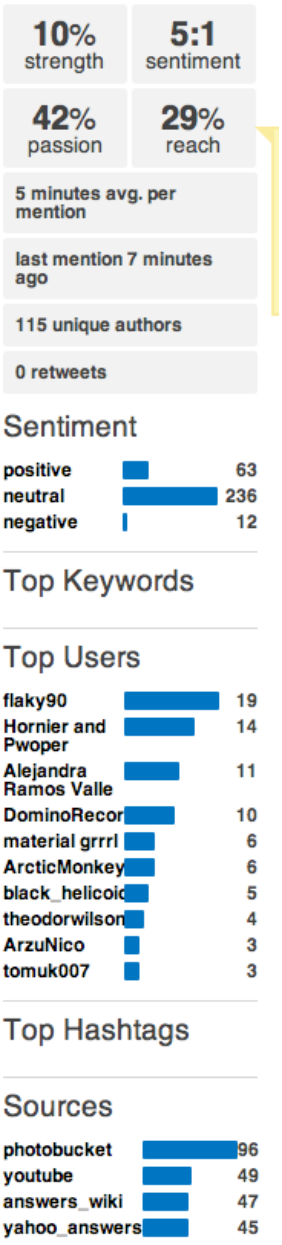
It should be noted that in most cases, it's not possible to know how the tools discussed here work. How transparent they are about their algorithms varies: some tools make concerted efforts to be transparent, but even with these, it's not possible to know precisely which websites and social media platforms have been included in searches and which have been left out. Choices about where to search for data and how to analyse found data will have shaped how all of the tools have been made, and these choices, in turn, shape the results. It is useful to bear this in mind when attempting to make sense of the data that such tools produce.

We thank Brandwatch (http://www.brandwatch.com/) and Meltwater Buzz (http://www.meltwater.com/) for making it possible for us to trial their platforms and services, and Mick Conroy for Tempero (http://tempero.co.uk/) for advice on our project.

1. Free online tools

1.1. Social Mention (http://www.socialmention.com/)





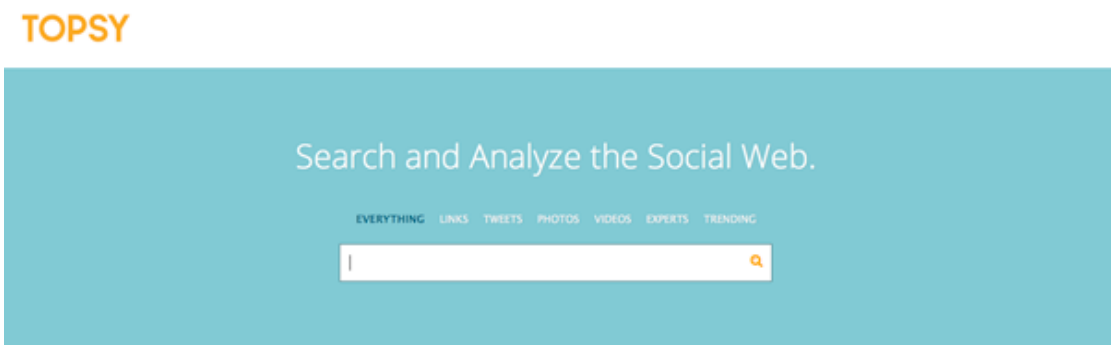
Social Mention is a free, online social media search and analysis platform that aggregates user-generated content from across social media sites.

At its simplest, it allows you to track and measure what people are saying about you or your company or any other topic across social media in real-time. Results from a simple search include sentiment, passion, reach, latest mentions, top users, hashtags and sources, as indicated in the image on the right.

Advanced services are also available.

1.2. Topsy (http://topsy.com/)

Topsy is similar to Social Mention, in that it offers a free, online search and analysis service. It divides its services into Social Search and Social Analytics. You can choose whether to search links, tweets, photos, videos or all of these, and you can compare and contrast mentions of a number of products or projects. More advanced services are offered through Topsy Pro Analytics.



1.3. Klout (http://klout.com/home), Kred (http://kred.com/), Peer Index (http://www.peerindex.com/)

These three platforms (Klout, Kred and Peer Index) measure and rate your interactions on social media sites in order to arrive at a social media influence ranking for you or your company. They all work in similar ways. The differences between them relate to: which sites they hook up to; how transparent they are about how they arrive at your score (a number of commentators suggest that Kred is the most transparent and Klout is the least transparent); and the extent to which they emphasise the perks of having a good score (these usually equate to discounts or gifts from brands to 'influential' people who, it is hoped, will review the products). Some people argue that the more transparent the algorithm, the easier it is for sites to be gamed.

1.3.1. Klout

Your influence is rated by the amount of involvement on several social sharing sites including Facebook, Twitter and more. Klout uses a different system to measure influence for each site available. For example, on Facebook, influence is measured by mentions, likes, comments, subscribers, wall posts and friends. On Twitter, influence is measured by retweets, mentions, list memberships, followers and replies. Once all of the factors for each website have been quantified, Klout takes the amount of reactions you generated on your social sites and compares it to the amount of content you share. This ratio then determines your finalized Klout Score.

1.3.2. Kred

Your rating on Kred falls into two categories, influence and outreach level. Influence measures how frequently and how many people make actions directly connected to your content on Twitter, Facebook and any other social media websites that you link to your Kred account. Kred focuses on how frequently you are retweeted, replied to, mentioned and followed on Twitter and the amount of posts, mentions, likes, shares and event

invitations to your name on Facebook. To calculate your outreach level, on Twitter, points are added when you retweet, reply or follow a new person and on Facebook, points are added when you post, like, mention or comment on someone else's Facebook content. Kred makes all this information transparent, meaning you are able to view each point given and how your points are transferred into your total score.

1.3.3. PeerIndex

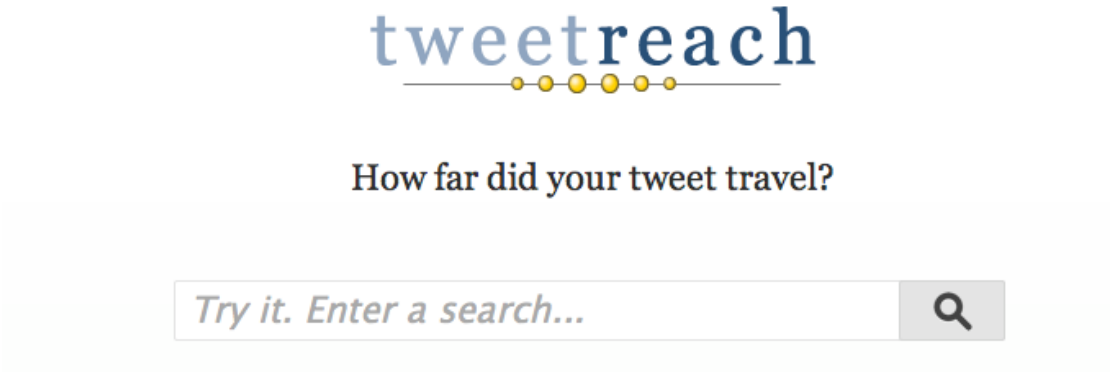
Your PeerIndex is measured by the content you create and the people who consume and react to it. As PeerIndex monitors the rate and quantity by which you share information, the site notes the type of content you are sharing and endorsing (by retweeting, liking etc.). As your content is interacted with online, your authority on a subject is increased. Some commentators suggest that PeerIndex is more focused on rewards and perks, or finding influential users who can review products.

For more detailed comparisions, see:

- 'Klout vs. Kred vs. PeerIndex: Reviews and Comparisons', http://thedsmgroup.com/klout-vs-kred-vs-peerindex-review/, 29/03/2013
- 'Don't Like Klout? 12 Other Ways to Track Social Media Influence & Engagement', http://blog.crazyegg.com/2013/06/04/dont-like-klout/, 06/04/2013

1.4. TweetReach (http://tweetreach.com/)

TweetReach is a free, online tool which allows you to identify the reach and exposure of specific tweets, hashtags, brand names, Twitter names or URLs and so gives an indication of the reach of Twitter campaigns. It reports on tweets and exposure, activity over time, top contributors and most retweeted tweets. Advanced services are also available.



1.5. SentiStrength (*http://sentistrength.wlv.ac.uk/*)

SentiStrength was developed by a computer scientist, Professor Mike Thelwall, to enable social scientists to use web and social media data in their research. It estimates the strength of positive and negative sentiment in online texts. SentiStrength reports *two* sentiment strengths:

-1 (not negative) to -5 (extremely negative)
1 (not positive) to 5 (extremely positive)

A free, Windows-only version of SentiStrength is downloadable, upon registration. A commercial version is also available. The tool can be tested out online.

Mike Thelwall has also produced a free online book, *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*, which is available at http://webometrics.wlv.ac.uk/.

1.6. Other online tools

HubSpot's Marketing Grader (http://marketing.grader.com/) assesses your website as well as social media reach, engagement and web optimization. It provides tips on how to build engagement and an assessment of social sharing of your content.

Sprout Social's #BePresent service (http://mustbepresent.com/#!home/) is a free service which allows organizations to get regular reports on how they are engaging on Twitter.

Other social media management tools such as **HootSuite** (https://hootsuite.com/) and **TweetDeck** (http://www.tweetdeck.com/) also offer analytics services, but often at a cost.

2. Digital Data Analysis project tools

Despite widespread claims, there is no such thing as 'getting the data' in the field of social media monitoring and analytics. Rather, *subsets* of data exist in accessible forms. Lots of different types of data exist, in different places and stored in different structures, with different access conditions. 140-character tweets with social network and profile data attached, Facebook status updates, likes and comments, blog posts, wiki edits, and web pages are all different things and access to each comes with unique usage limits, constraints and difficulties. The level of access often changes depending upon how much one pays, or which deals have been brokered to exploit data sources for commercial gain. Additionally, access to this data can take many forms, such as exports of raw data or of transformed data, provided in useful formats such as social network maps or automatically created content clouds. Data sources must be chosen carefully in order to make sure that they are appropriate for the project. Below is a short introduction to some of the free or affordable tools that exist that allow, with a bit of manual work and expertise, the investigation of online data.

2.1. DataSift (http://datasift.com)

DataSift is a commercial, online tool that harvests live data from a variety of social media as well as a comprehensive cache of past contributions. It is free to sign up to and you get $10-worth of data for free. This is a lot of data. The cost of data streams is per Data Processing Unit (DPU) and 1 DPU for 1 hour costs $0.20. A simple process uses less than 0.5 DPU. Theoretically you can build queries that access social networks in real time and via a cache to harvest content about a particular term or account. The data sets produced can then be analysed with other tools discussed in subsequent sections.

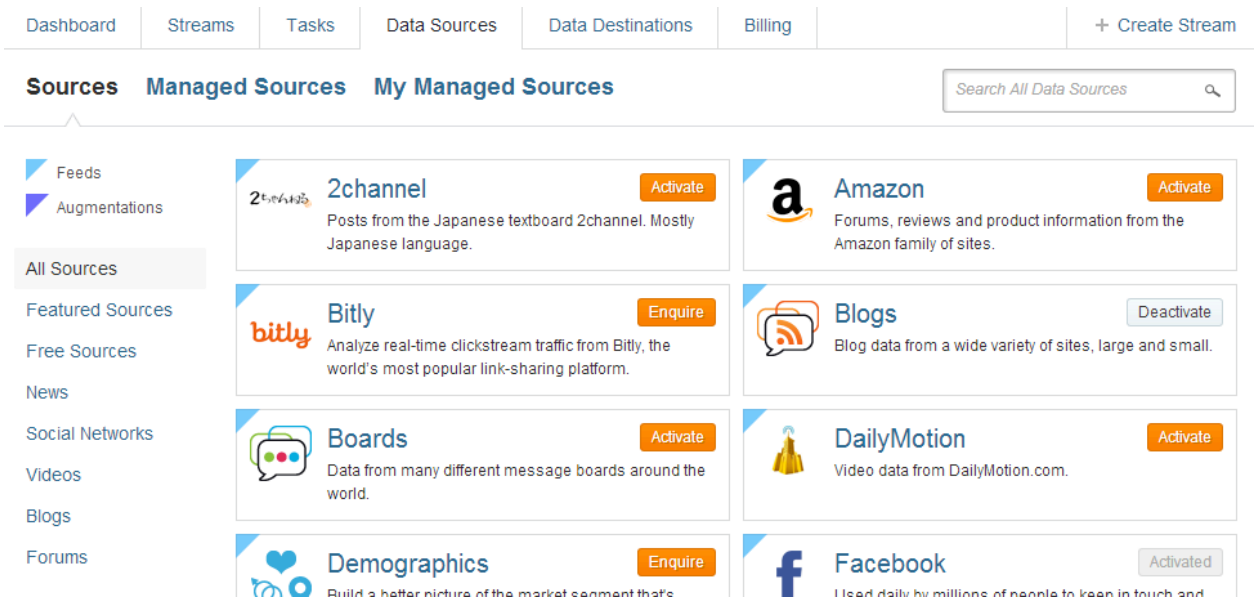
The user can specify search terms and search durations, then watch as the data rolls in. The result sets created can be exported in CSV format – effectively one row in a spreadsheet for each piece of data found. DataSift calls one record of online activity an "interaction". Every tweet is an interaction; every Facebook post is an interaction, and so is every blog post, forum contribution, etc. The DataSift output file contains a record of all relevant interactions found by your search. Each row in the output represents an interaction, each column a field.

DataSift is a commercial service and targets most commercially valuable data sources. It accesses a very large percentage of the sources of online data, but those are not necessarily the specific sources that are valuable in searches of niche environments. For instance, blogging sites such as *Blogger* and *WordPress* are indexed well, as are large forum sites such as *Topix*, but local forum sites such as *LeedsForum* and *SheffieldForum* often do not appear in results sets, presumably because they are not as commercially important so have not been targeted.

Selecting Data Sources

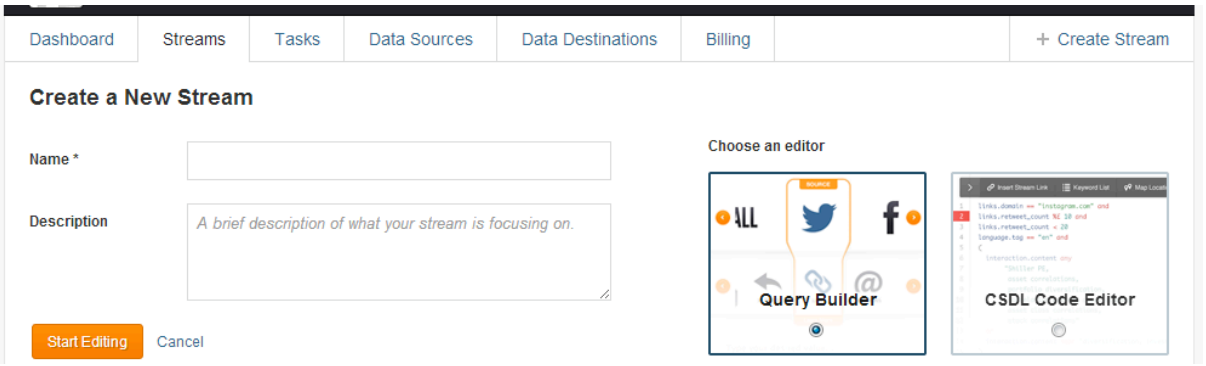
DataSift can look at lots of different online platforms when searching for data, such as Facebook, Twitter, blogs, forums, Wikipedia, and many more. These come at varying

costs. Facebook data comes at no extra charge, as do a few other sources. Twitter comes with a very small charge and other data sources get increasingly expensive (blogs cost $2 per 1000 interactions). After logging in the user can navigate to the 'Data Sources' tab to view the whole list of available sources, free and paid for. The list includes 'augmentations' – these are DataSift add-ons that generate extra data from the search results using bespoke algorithms. 'Demographics', for instance, attempts to find the gender, location and other data about each interaction by accessing all of the information available across the different platforms.

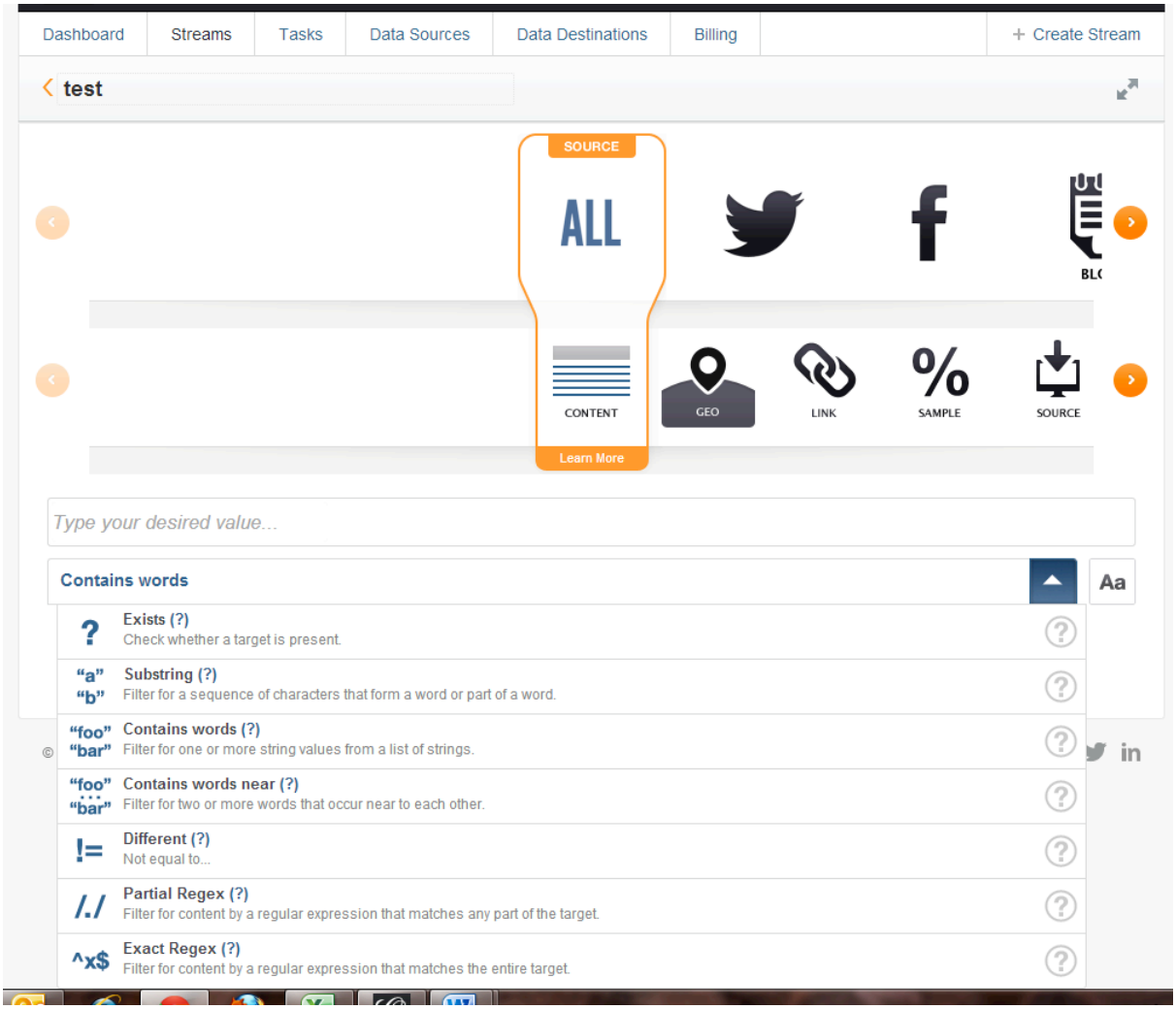


Creating Data Streams

After logging into DataSift, click on the Streams tab and then 'create new stream' to get the following screen:

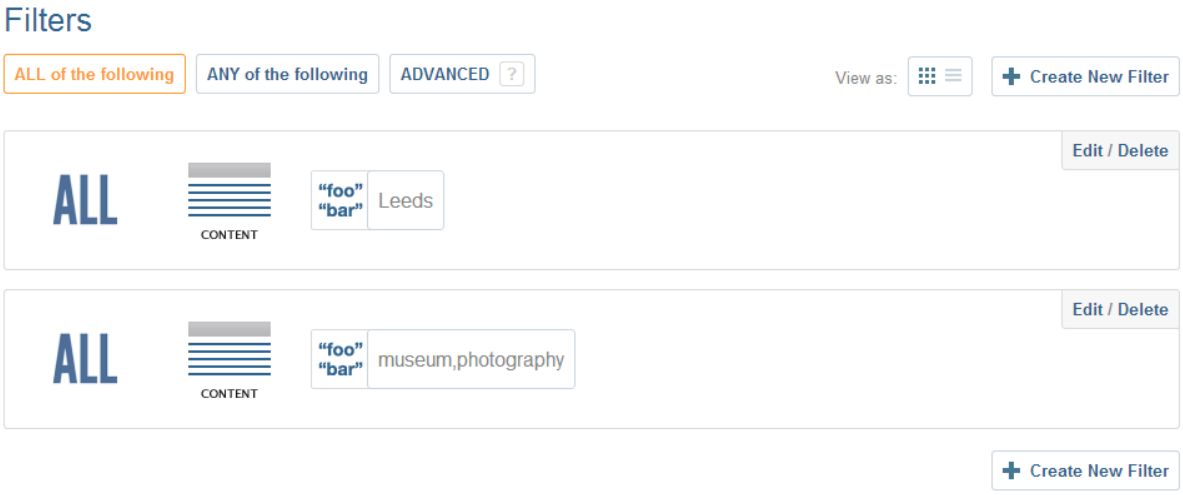


Add a name and description then click 'Start Editing' to create your search. Add a new filter:

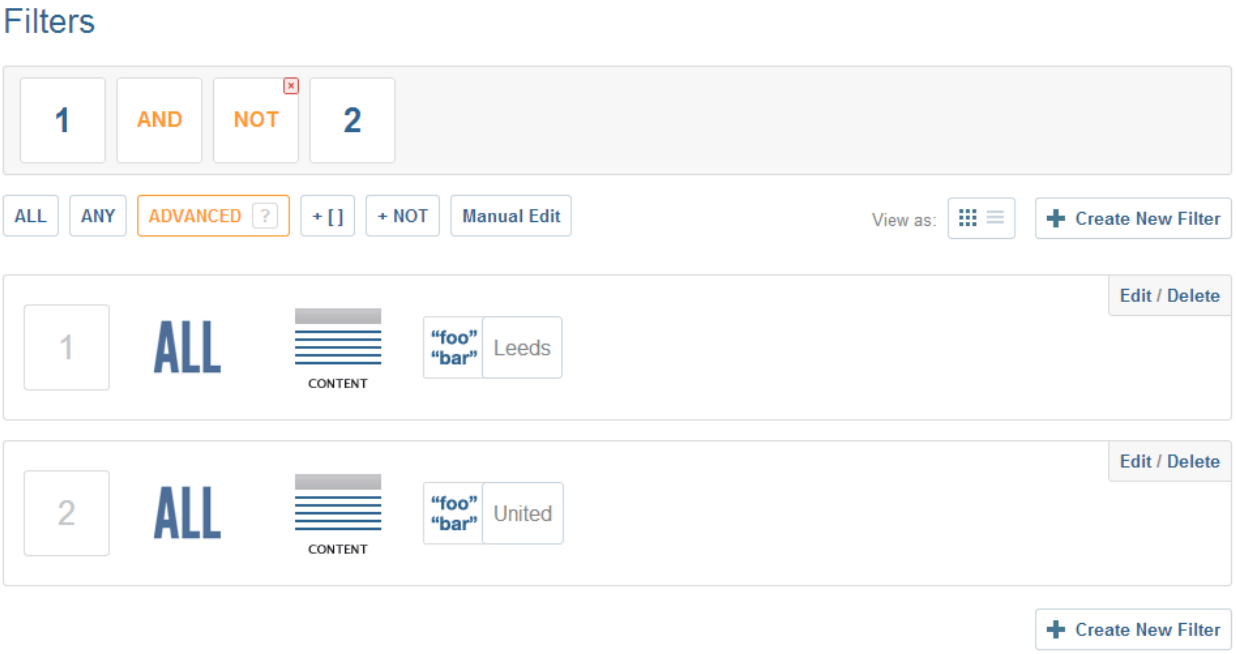


The most common filter to add is a text filter. You can choose to filter only in specific platforms (the top row of the screenshot above) or on any component of that platform (the second row). Typically you will select 'all' and 'content' to filter just relevant content. You can choose the type of search from the list in the screenshot above. 'Contains words' (= search for content containing any of the terms entered) is the most common, but there are other useful options such as specifying that words must appear near to one another (within 15 words is often considered as being in the same sentence). You can add more than one filter too, to require more than one term. For instance, to search for 'museum' or 'photography' AND 'Leeds' create two filters, one looking for 'Leeds' and the other looking for either of 'museum' or 'photography'. Just separate these with a comma when typing them in.

The stream will look like this after the filters are added:



Notice the three tabs above the filters – these allow increased flexibility when searching, allowing the user to select whether ALL filters have to be met to return data or just one of them. In the example above, using the ALL option would mean that content would have to contain the word 'Leeds' as well as 'museum' or 'photography'. Using the ANY option would mean that content with at least one of those words would be returned in the results. The ADVANCED option allows more complicated searches to be done, for example requiring multiple filters in specific ways or using the NOT keyword. This keyword allows the user to specify content that should be excluded from the result set, for instance the word 'United' to remove football related content from a search:

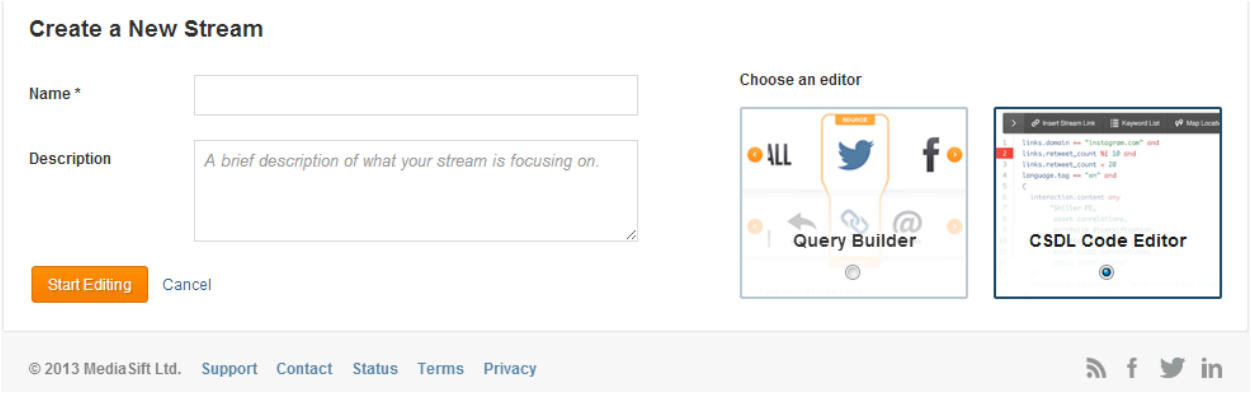


DataSift can provide affordable access to large quantities of data, but due to its global reach, the challenge is to find specific local data. Careful keyword selection is vital, as well as intelligent search filters that look at location (as far as possible) and sentence context. Very specific keywords (such as jargon terms) often fail to harvest content, while others are too general and yield thousands of irrelevant results (words like 'market' that can be used in lots of different contexts). Therefore a careful process of

combination, refinement and trial and error needs to take place to find meaningful data sets.

Further to this process, it might be necessary to restrict content to that produced in a particular location, filtering out data from the US, for example. Unfortunately, geographical data is not as common or easy to analyse in social media as is often assumed. Platforms such as Twitter do make it possible to attach latitude and longitude values to contributions, showing exactly where the user was when they wrote the contribution, but these are often not used. A user has to be using a device that supports geolocation, turn on geolocation on the device and then give the platform permission to use that geolocation data. Other platforms do not support geolocation data at all. Other location data is available, such as the location or time zone fields in a user profile but the user is usually free to leave these fields blank, or even to put in false information. Because of these factors, typically around 10% of data have valid geolocation information and a few more have timezone information attached. Depending upon the platform used, a large proportion of relevant contributions can be lost when a locational filter is applied.

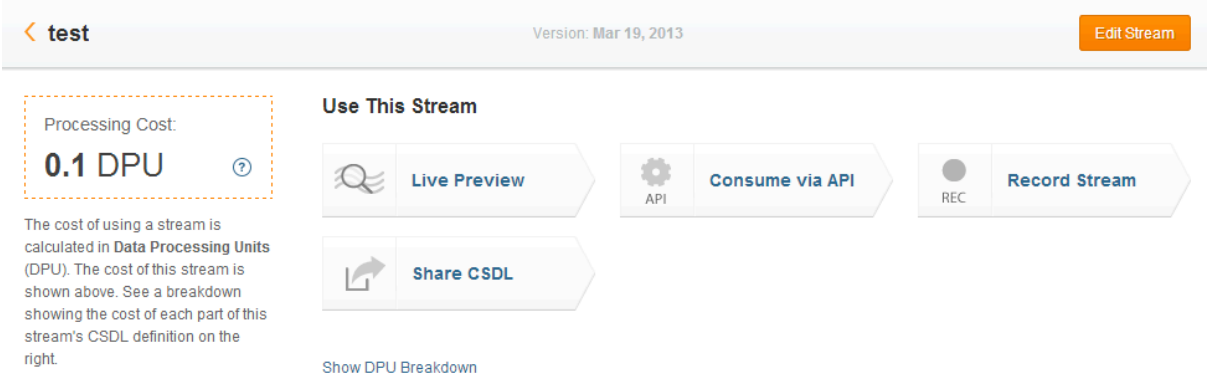
Using the 'CDSL' editor when creating a query, it is possible to utilise a locational filter that combines a range of locational data, such as geolocation coordinates, Twitter profile data and time zone data, and using the OR keyword to minimise exclusion (see the code in the image below). However, this technique will still exclude relevant contributions if they do not contain any location data – which will be a large number if the search is done on multiple platforms. For this reason, finding keywords that accurately match content to your topics is a much better way to maximise the accuracy of your search results than using locational filters.



Once the stream has been defined, you will see options to utilise it. The most important options are *Live Preview* (watch real time occurrences of contributions as people write them – this costs nothing) and to *Record* (collect real-time and contributions from the recent past – this option uses up credit and provides data that can be exported). These two options are highlighted in the next sections of this report.

Previewing Data Streams

Harvesting data can be a long and expensive process and when you get your search wrong, it can be wasteful and annoying! A really good thing to do first is to preview your search results using the 'Live Preview' button:

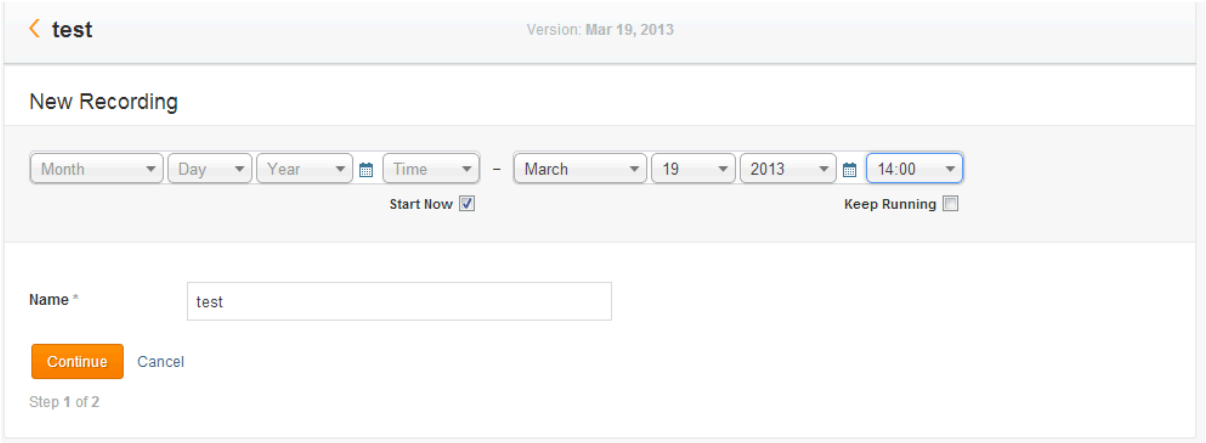


This will open a screen that shows the content that is found by DataSift as it runs your search, but it does not store the data. This costs much less, and allows you to view the data and decide whether it is relevant and useful. If it isn't, you have not wasted much time or money and can edit the stream. If it is, you can stop the preview and record the stream.

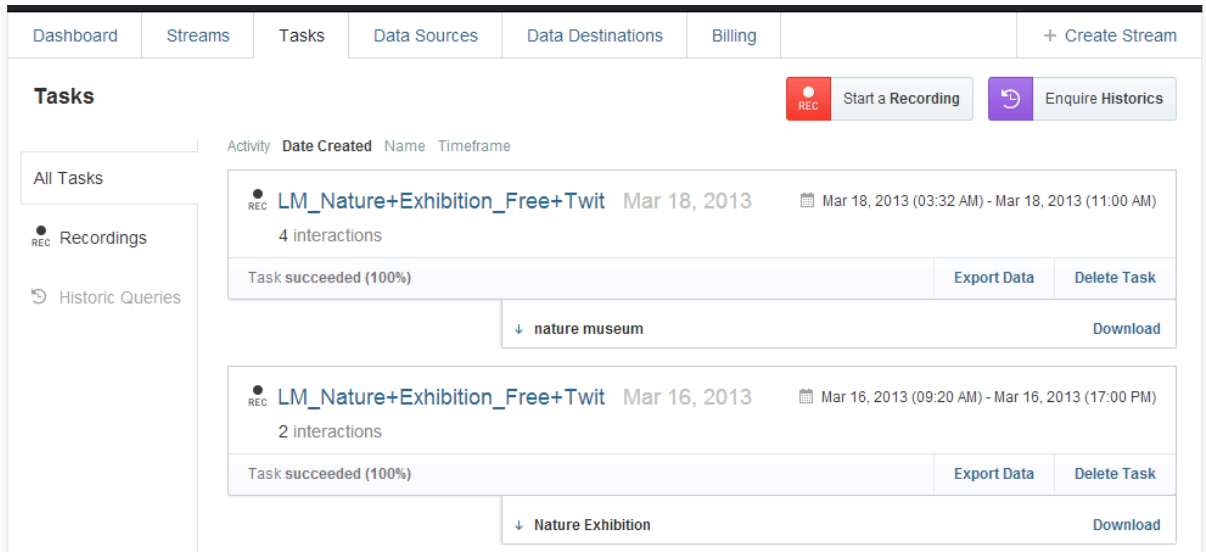
Recording Data Streams

Recording the streams that you have created stores the data on the DataSift server, from which you can download it. Recording costs money, so make sure to only record streams that you have tested in 'Live Preview' (see above). After you click 'Record Stream' you can define the length of time the stream should be recorded for. It seems sensible to

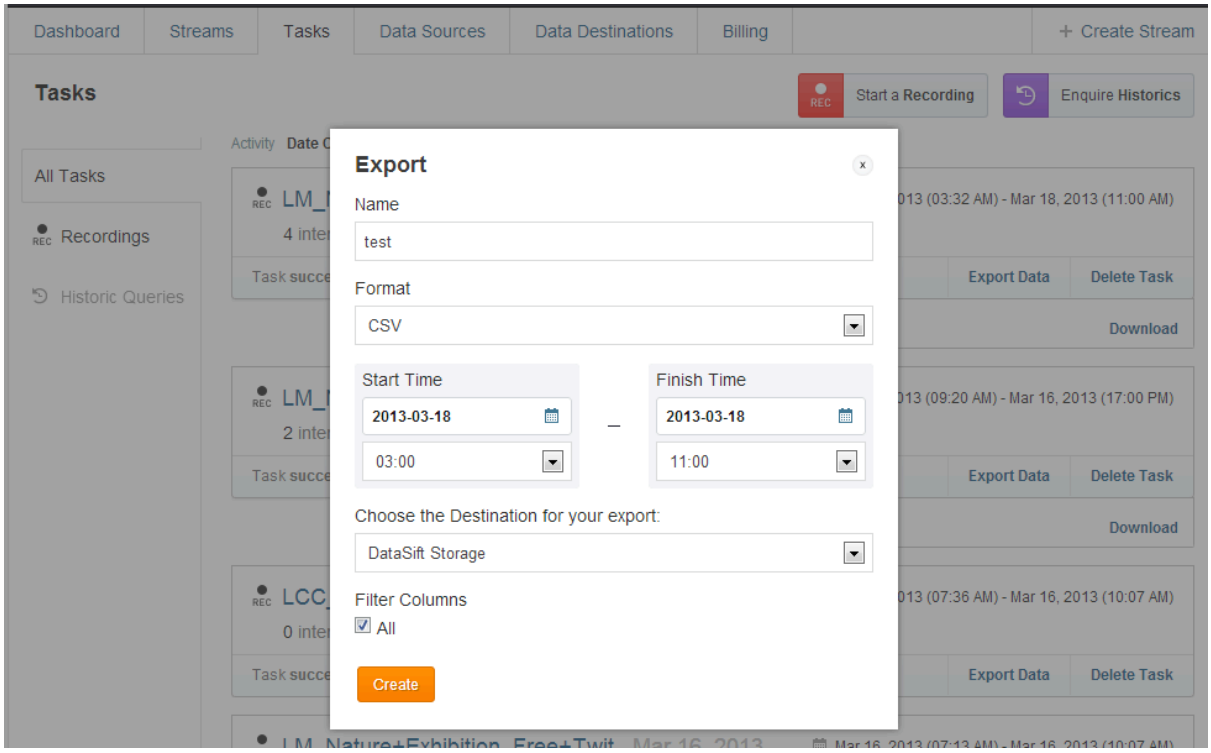
record for a small amount of time first, but unfortunately a quirk of the system makes this difficult. The DataSift server runs on California time by default, and you need to make sure that your profile time zone is set to UK so that the recording process makes sense! Luckily you can stop a recording at any time, so if it is using up all of your credit, you don't have to leave it running.



Once you have set a stream to record, a summary of the recording will appear in the 'Tasks' section, showing the progress of the recording, the number of interactions harvested and allowing you to export and download the data:



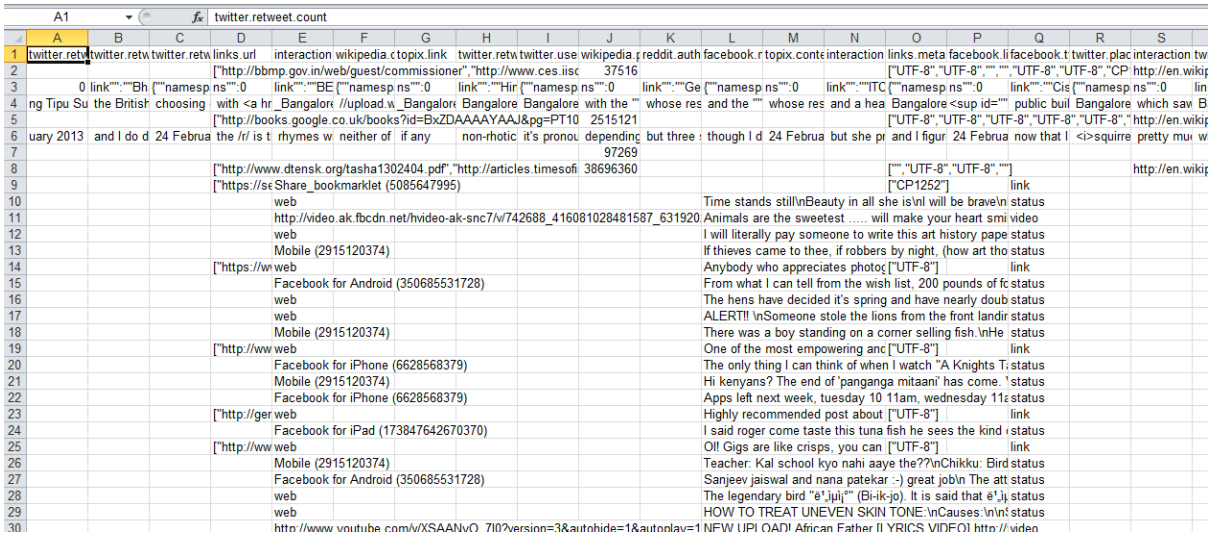
After clicking on 'Export the Data', you can define the type of export (usually CSV) and set it running. This might take a long time for large data sets.



When the process finishes you can download the file by clicking on the 'Download' link that appears in the tasks screen.

Output Data

DataSift outputs have many fields (columns) of data – one for each component of each data source activated (i.e. one field for each component of Twitter data available, one for each Facebook component, one for each Wikipedia component, and so on) plus extra columns of amalgamated 'augmentation' data that DataSift produces to make the output more useful. For instance, the 'interaction.content' field contains the main content of a record – whether that is from a Facebook post or a Tweet or a blog post. Thus, every interaction (row) has a 'content' field that contains the important contribution data. These fields are present for every record (every 'interaction'), even if they are not applicable – interactions from Facebook still have all the Twitter fields, they are just blank:



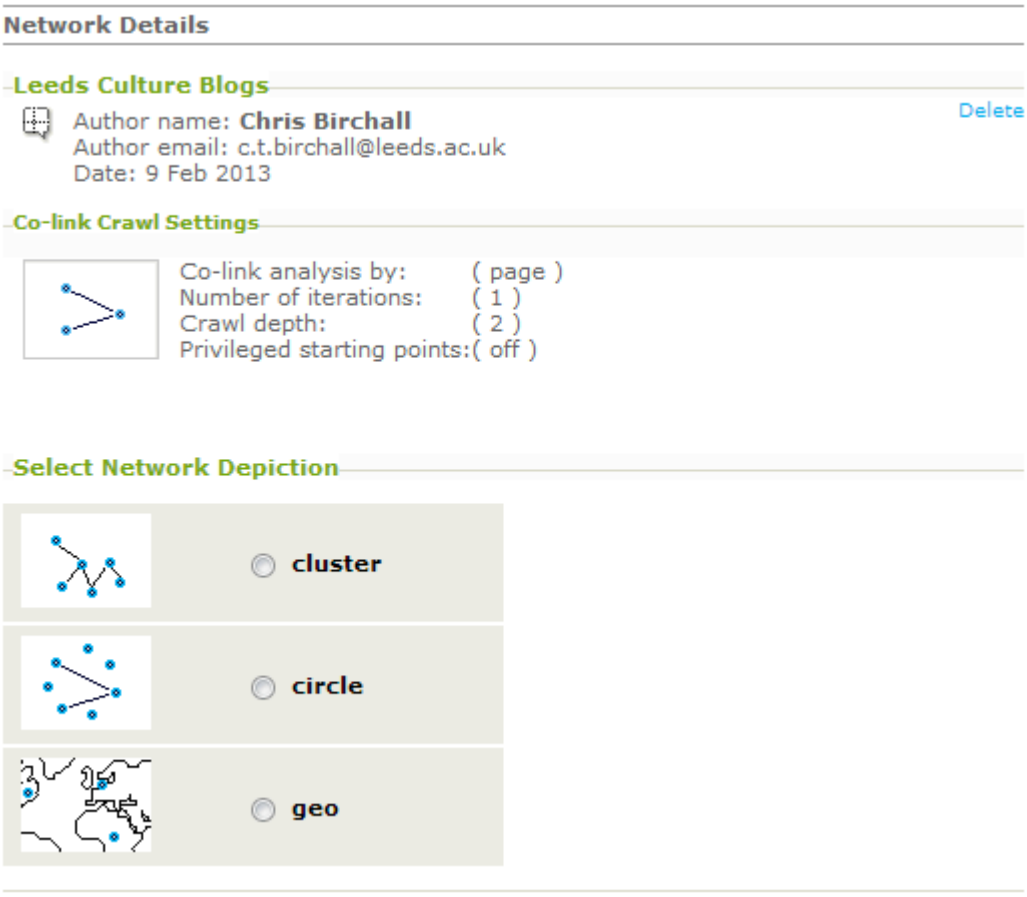
In this CSV data export (opened in MS Excel), the top row contains field names and the rest the data for individual interactions. You have to do some exploring to find the data that you want, but you can do a 'find' operation to get to the 'interaction.content field', where the main post content resides. There are other columns such as 'twitter.retweet.user.name' and 'twitter.retweeted.user.name' that identify linked contributors. Using these columns in network analysis is possible, and is detailed briefly in the NodeXL section of this report.

2.2. IssueCrawler (www.issuecrawler.net)

IssueCrawler is an academic tool for discovering linked content on the web. Users can sign up for free but accounts must be approved by the site administrators. Academic researchers are usually approved, whereas commercial users may not be approved automatically. It may be possible for employees from non-commercial organisations to sign up to use the tool. The free account allows up to ten searches before the user is asked to contribute. There is no published payment scheme.

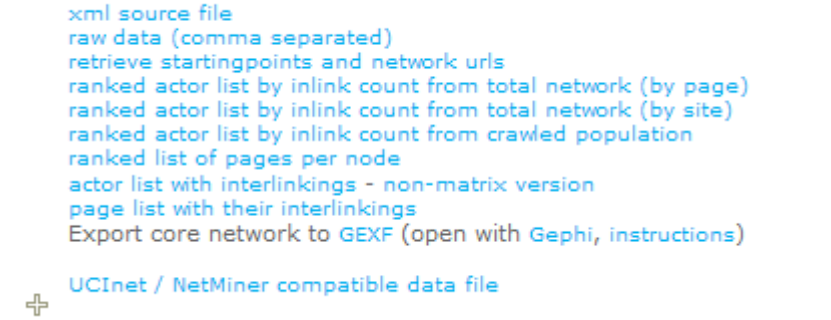
Like some of the other tools in this document, IssueCrawler relies upon some prior knowledge to ensure high quality results. Searches must be 'seeded' with relevant URLs – known websites that have content that is relevant to a required search. These might be known blogs or portals that may be identified by prior knowledge, simple web searches using tools such as Google, or by analysis of social networks using tools such as NodeXL (discussed below). Most likely, a combination of these will be used to build a comprehensive list of known content. The list of URLS is entered into the box on the 'Issue Crawler' tab after the user logs in:

After seeding the search, the user configures it, choosing a type of analysis, number of iterations to carry out and the crawl depth (other characteristics can be set, but are generally just left at the default values). There are three types of analysis: co-link, snowball and interactive. Co-link analysis is the most useful in our context, as it only includes sites in the network that have two links to them from our seed list, generating communities of content that contain sites that link strongly to each other. The iteration value controls how many times this co-link analysis is done. A value of 1 means that the seed sites are loaded and co-link analysis is performed once. A value of 2 means that the links in the external pages discovered in the first iteration are loaded and a new co-link analysis is performed. The depth value can amplify this effect by causing more than one level of links to be followed, and pages loaded, between each iteration, making the search much wider. For the most in-depth results, change the number of iterations to three and a crawl depth of two (looks at the links within seeded pages and links in pages that they link to):



IssueCrawler then follows the links within these seed pages and builds up a picture of connected content on the internet. It can help to find new, previously unknown blogs and websites that publish information about a topic of interest and also identifies how each of these sources link to one another, building up a picture of content pathways and influence that affect the user experience as they consume online content related to your topic.

The crawl is queued for processing, usually taking a few days to complete. When it has completed, it becomes available for download. The website will display a list of the sites found or allows the download of a map file that can be opened in visualisation tools such as Gephi (the section on Gephi, below, details how to visualise and analyse these maps). The various options for viewing results and downloads are available as a list, underneath the details of the crawl on the website.



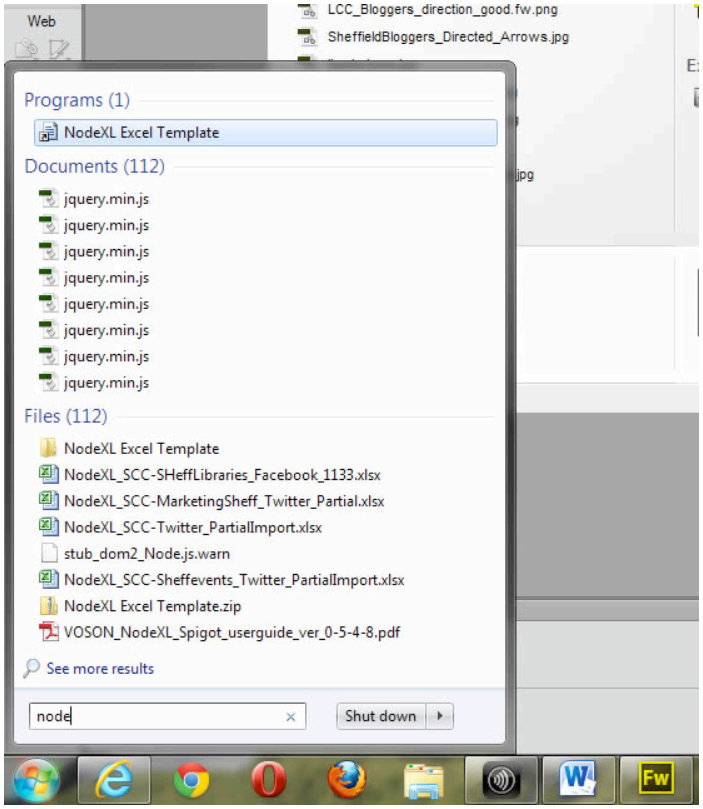
Manually verified data sets, such as lists of bloggers, can be fed into other social network analysis tools, such as ShareCounter, a tool that will run through the list of identified blogs to find out how much their content has been shared on a range of social networks, such as Tumblr and YouTube. Combined with a Gephi visualisation, it is possible to build up a picture of influential content producers and distributors that are important in your engagement with the public.

2.3. NodeXL (*http://nodexl.codeplex.com/*)

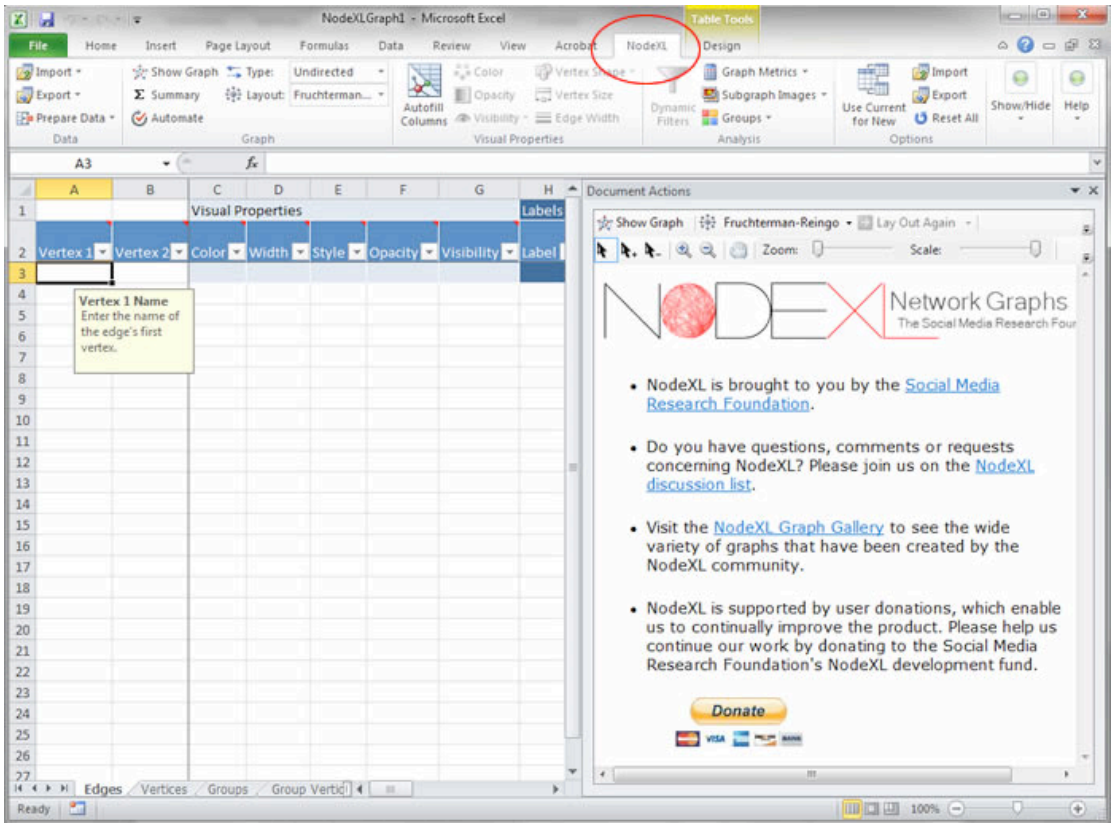
NodeXL (and its associated Social Media Importer plugin) is a freely available download from the Social Media Research Foundation that adds both network analysis and social media connectivity capabilities to the Microsoft Office Excel program. It runs on Windows machines only.

Using NodeXL you can harvest data from a variety of sources (Twitter, YouTube, Flickr, email, Facebook and WWW hyperlinks) about people/organisations that are connected to your accounts. However, unlike other data harvesting tools, such as DataSift, NodeXL can only access one of these sources at a time, so you can build up networks from Twitter data or from Facebook data, but not from both together. It can take several hours to analyse a large network, due to usage factors such as the Twitter API request limit.

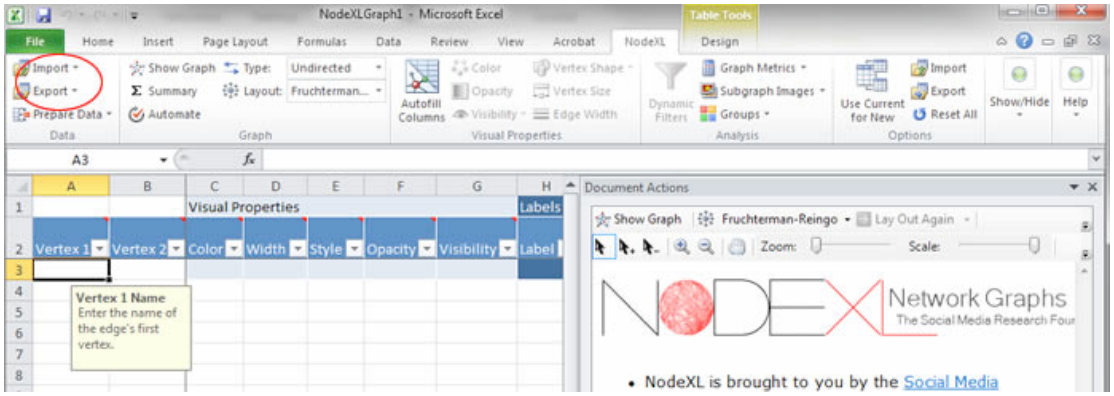
NodeXL can be difficult to find at first, as it is an excel template file, not a standalone program. The easiest way is to type 'NodeXL' into the Windows search pane and select 'NodeXL Excel Template':



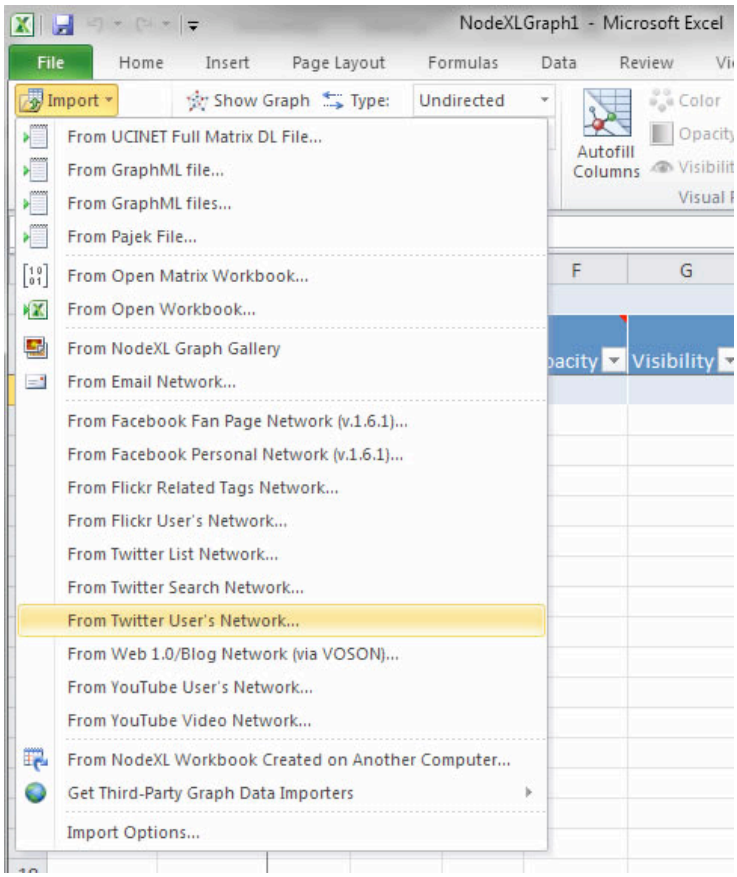
Microsoft Excel will load, but there will be column headers already defined and maybe a graph panel visible. There will also be an extra tab on the ribbon, entitled NodeXL. This is where all the extra functionality can be found:



On the left hand side of this new tab are drop-down boxes entitled 'import' and 'export'. These are the most important parts of this screen as they let us harvest data from social networks and save it in a variety of formats for later use.

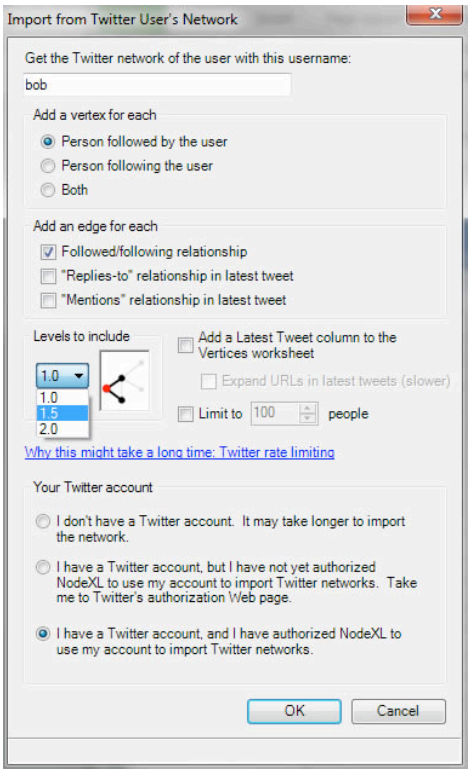


Clicking on the import button brings up a dialog box that allows us to select a data source from a whole host of online platforms; perhaps the most valuable of these are the Twitter Users network (to create maps of people linked to a Twitter account), Twitter Search Network (to create maps of people that have used a keyword in their Tweets) and Facebook Fan Page (to create maps of people that have interacted on a Facebook Fan Page):



Analysing a Twitter User Network

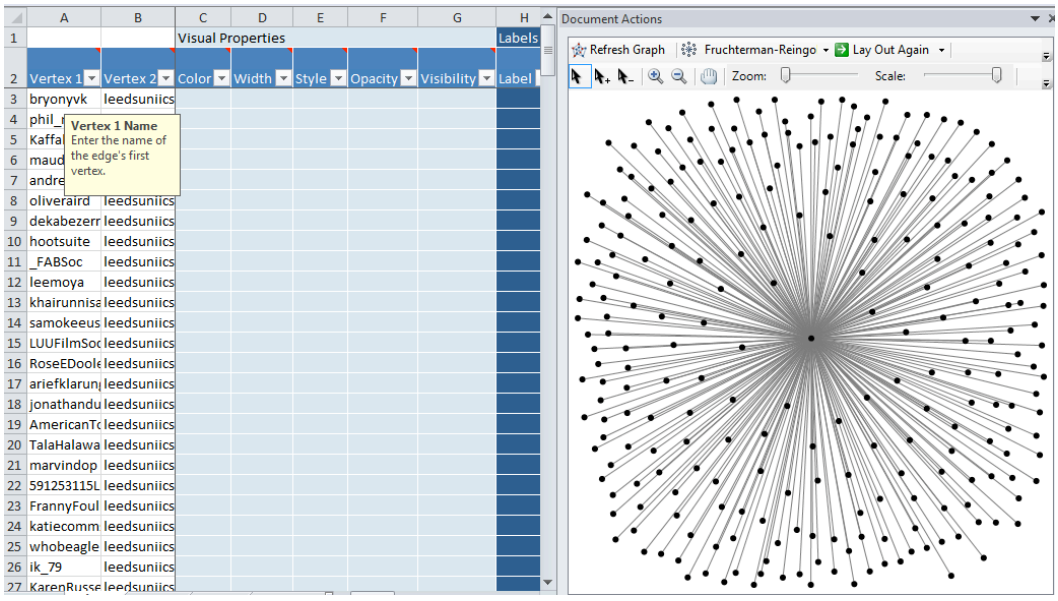
Selecting Twitter User's Network allows us to configure the search that is performed. By entering the name of an account and selecting the type of relationship that we want to include (followers, replies, mentions, etc), the scope of the network to be produced can be controlled. There is also the option of 'levels to include'; this determines whether NodeXL finds just the followers of an account (level 1), the followers of the account plus whether any of them follow each other (level 1.5), or all the followers and all their followers (level 2).



You have to have a Twitter account in order to search the database and you must allow NodeXL to use that account. This allows Twitter to keep track of who is using its data.

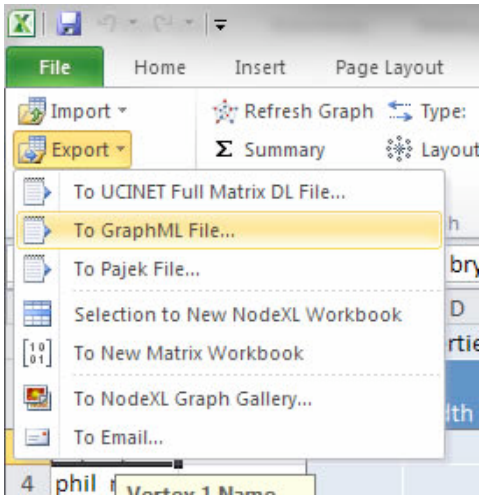
Level 1 searches are rarely very interesting – they simply show how many followers an account has (though could be useful if you are defining a relationship as someone who mentions the account or replies to it). Level 2 searches usually get very big and take a long time to complete due to the Twitter rate limits (free tools like this can only make 150-400 requests to the database per hour). Level 1.5 searches are therefore often the best way to go.

After filling in the details and clicking OK, the search will begin. When it completes, the first two columns of the spreadsheet will fill with the IDs of connected Twitter accounts. Clicking 'Refresh Graph' in the graph pane displays your newly discovered network:



To make full use of this data, we can alter the map using the built in visualisation and statistical tools (all are listed above the graph within NodeXL) but we can do this much more easily in more powerful visualisation tools such as Gephi (see section below). In

order to view the data in other programs, the export function of NodeXL can be used – upon clicking this button a variety of export formats can be selected:

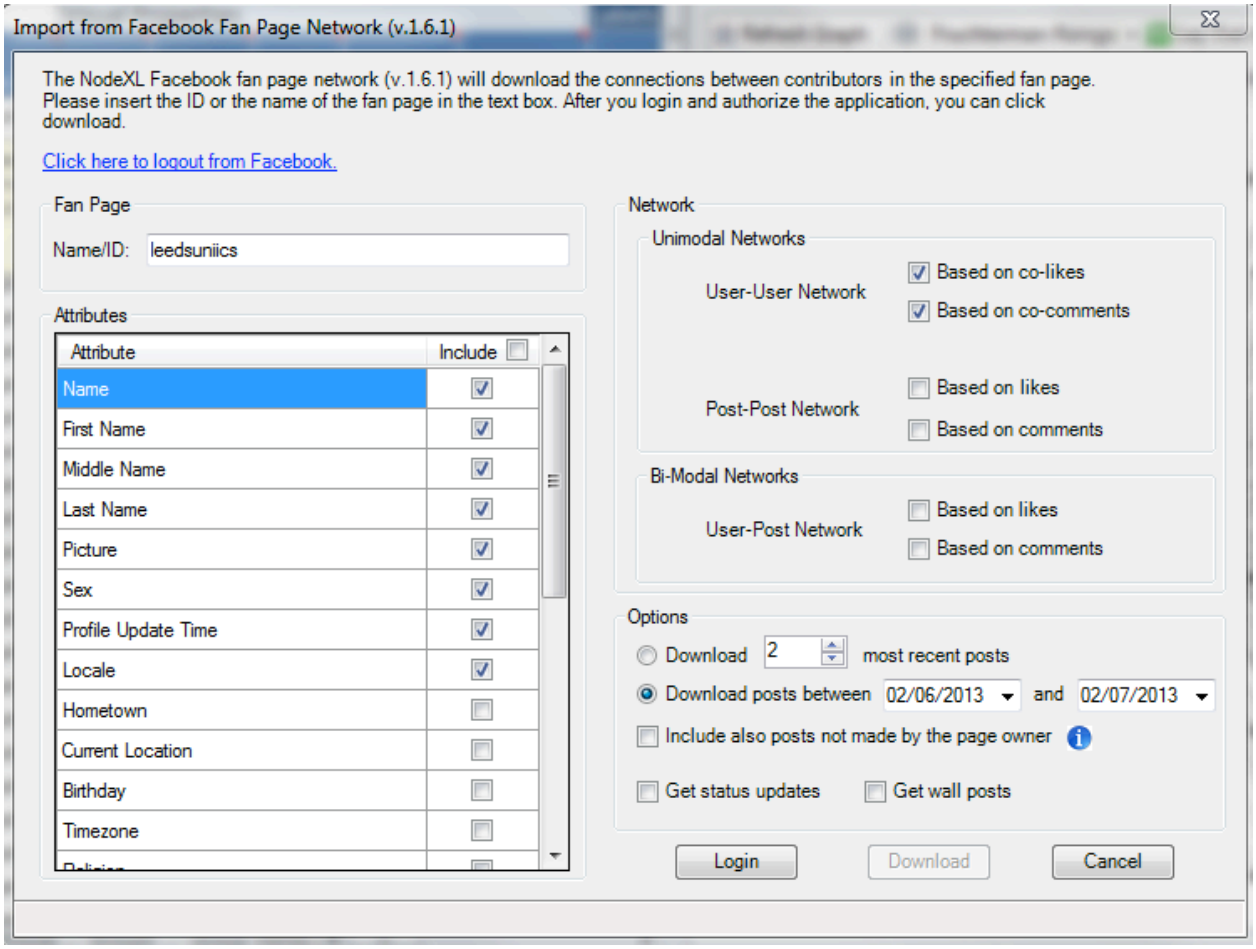


The GraphML format is very useful – it creates small, lightweight files that can be opened and edited in a number of visualisation and network analysis tools, including Gephi.

As described below in the section on Gephi, different Twitter accounts can be mapped to show how well connected they are – not just through number of followers, but also taking into account the number of users to whom each follower is connected. Influential contributors and communities of contributors can be identified. These maps can be used in order to target information at particular accounts, and the different social media accounts owned by one organisation can be compared.

Analysing a Facebook Fan Page Network
Using NodeXL, Facebook accounts and Fan Pages can be analysed in a similar way to Twitter accounts. Very popular fan pages generate lots of interactions between fans; most other organisations' pages have fewer. There are several types of connections identified within Facebook contributions: unimodal connections such as user-user (users that like the same thing, or comment on the same thing); post-post (content that is liked, or commented on by the same people); and bimodal connections such as user-post (which content a user has liked or commented on). The most important of these might be the user-user, but we can also look at post-post connections, if we are interested in which pieces of content stimulated connections.

To harvest Facebook fan page data, select 'Facebook fan page' from the import menu and fill in the dialog that appears, including the types of data that you want to collect, types of interactions in which you are interested and the date range that you want to sample:



Again, you have to log in to the target service to enable NodeXL to access the data, but once you have logged in you can click download and the details roll in as they did with Twitter. You can save or export in the same way.

Getting Data From DataSift

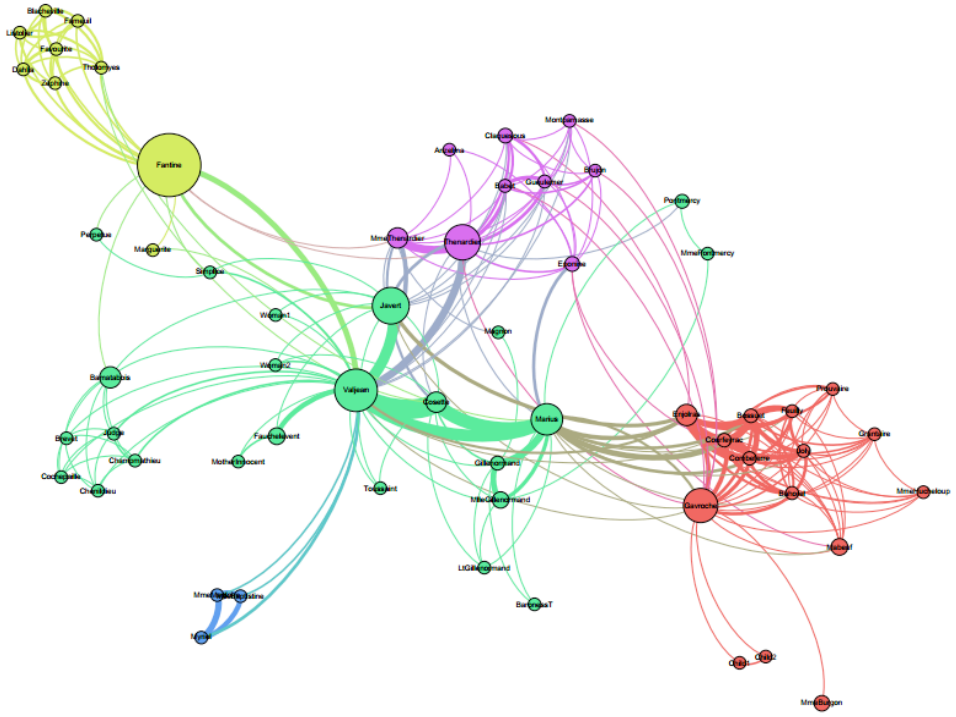
The data exported from DataSift (detailed earlier) contains information about social connections between contributors. For instance, if an interaction in the data set is a retweet, the row of data will include the username of the retweeter and the username of the original tweeter. These can be extracted, with a bit of Excel manipulation, and copied into NodeXL to form a network which can then be analysed. To find all retweets in a data set, perform the following steps:

- Find the twitter.retweet.user.name and twitter.retweeted.user.name columns in the DataSift export
- Copy these columns into a blank worksheet
- Filter the worksheet so that there are no blank lines in either column (and check for Excel errors such as "?NAME" which can creep in when content from DataSift looks like an Excel formula). Retain only rows with valid data in each column.
- Copy this new, filtered data from the worksheet and paste into the vertices column in NodeXL
- Export to GraphML, for efficient storage and use in other programs such as Gephi and Pajek.

Finding data in this way takes a lot of manual effort. It can be really valuable, if you have got a very accurate and relevant DataSift export, but might be a waste of time if the original search has not produced relevant data.

NodeXL can perform analysis based upon a number of metrics, such as connectedness, betweenness centrality (influencers) and community identification. This process is even easier if the data is imported into Gephi for analysis. The key influencers and communities identified can be analysed; hyperlinks can be found that can be used in subsequent analysis (below) and outreach teams can communicate with influential groups and individuals to control information flow.

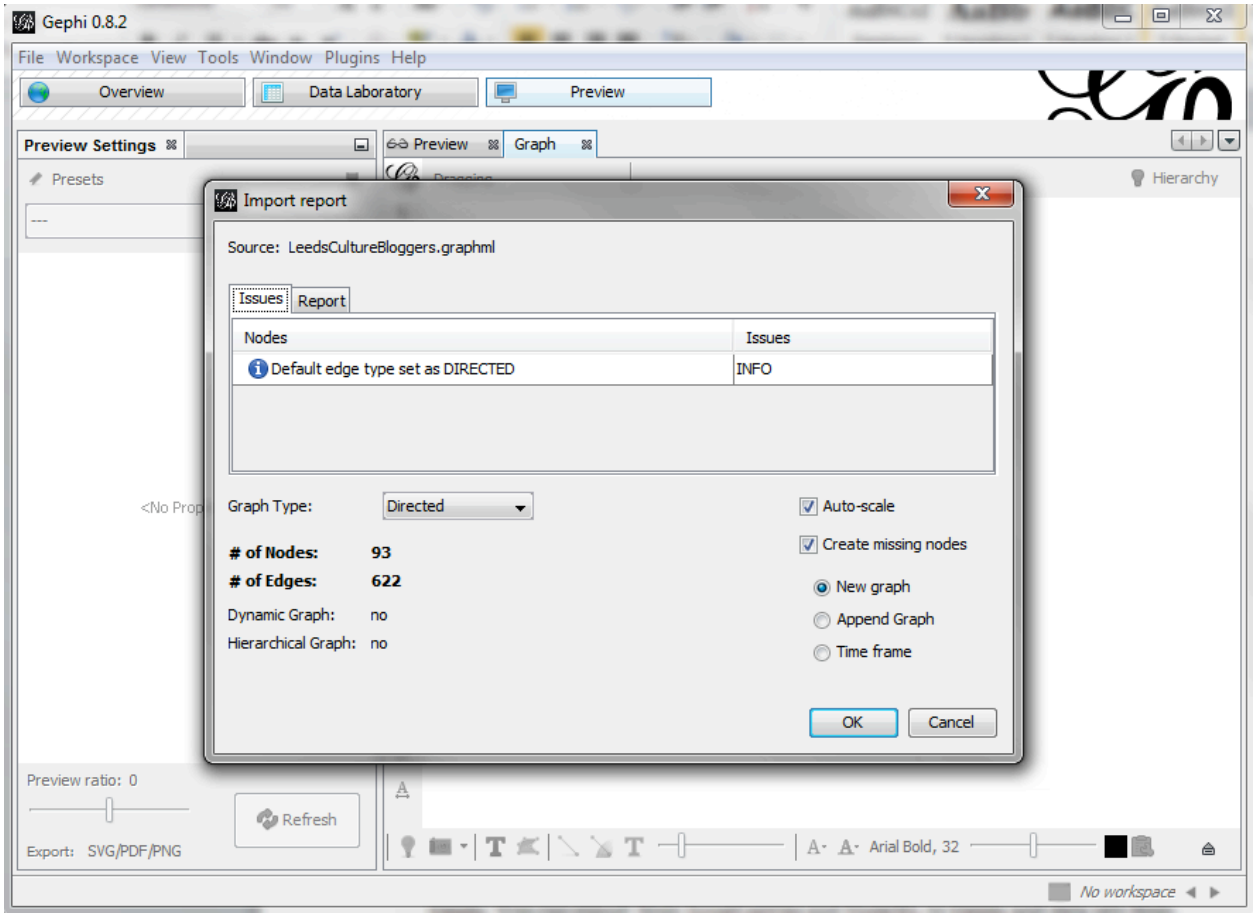
The diagram below shows a conversation, formatted to show the most influential nodes, or gatekeepers, (the large circles) as well as potential communities (the coloured groupings). The links between nodes could represent hyperlinks between websites in some data sets, or links between participants in a conversation in another data set. The next section of this report details how to analyse and visualise this raw network data to get meaningful information about account reach, communities of followers and influential followers.



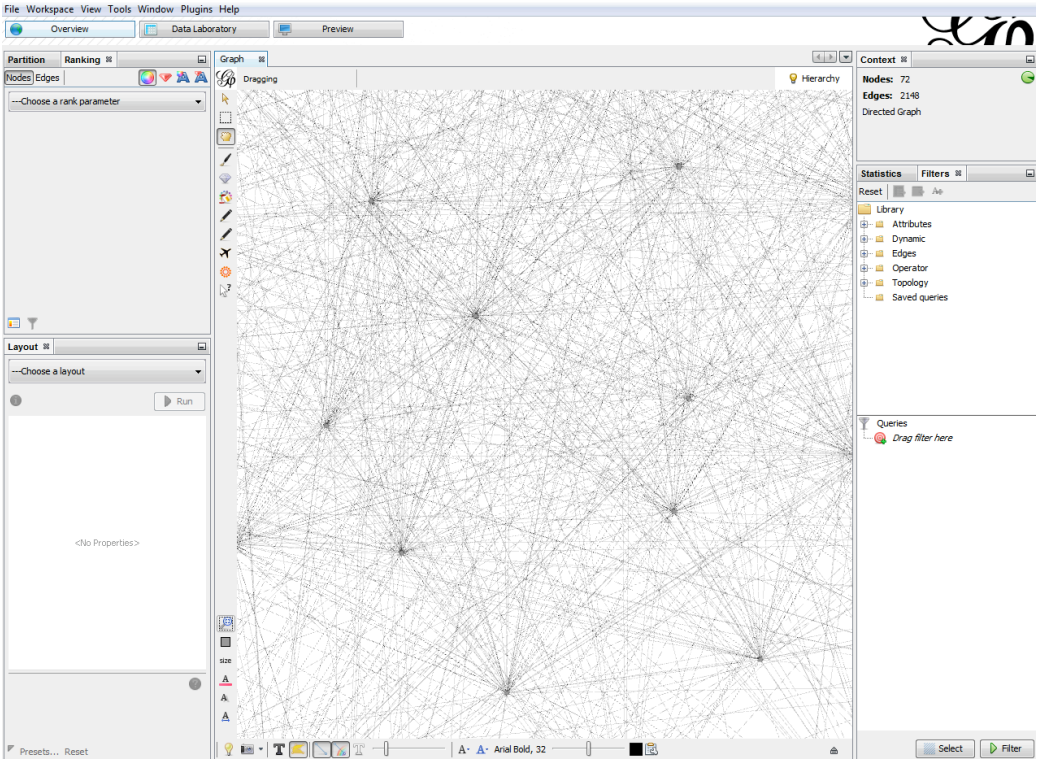
2.4. Gephi (https://gephi.org/)

Gephi is standalone software that can be installed on any PC/Mac. Gephi allows very easy graphical representation of the 'connectedness' (degree), 'influence' (betweenness centrality) and community membership of individuals within a network. It can do a lot more, but these are the most valuable initial analysis steps that provide insight into the structure of our social networks. Gephi does not have plug-ins that harvest data directly from social networks. Instead, data sets need to be produced in tools such as NodeXL and then imported into Gephi. You can export from IssueCrawler and NodeXL to

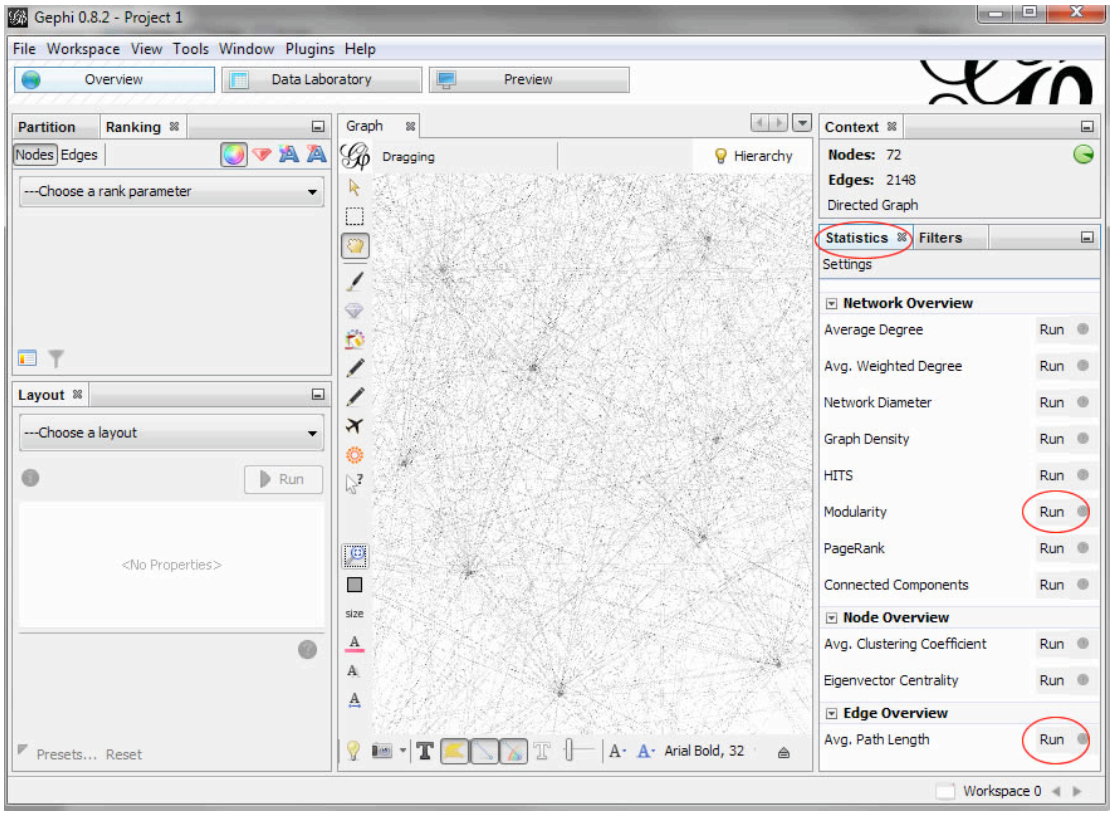
Gephi and data sets from other services can also be compatible. To open a data set exported as a GraphML file from NodeXL, simply open Gephi and open the GraphML file by clicking on File and Open from the menu bar. A dialog box appears, asking what you want to do with the data:



Simply accept the default options and click OK. The map will appear in the Gephi window.



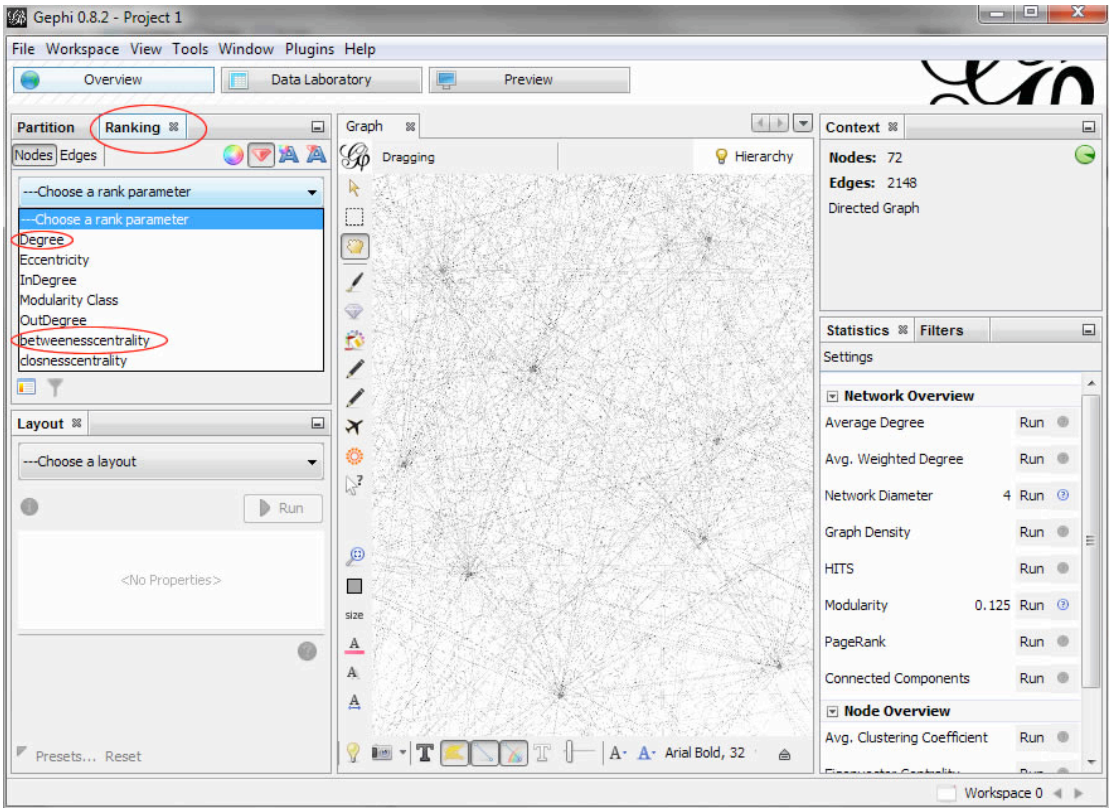
There are now several steps to be followed to turn this randomly displayed map into something more useful. The first step is to run a few statistics. Find the Statistics panel on the right hand side and run the 'Ave. Path Length' test and the 'Modularity' test:



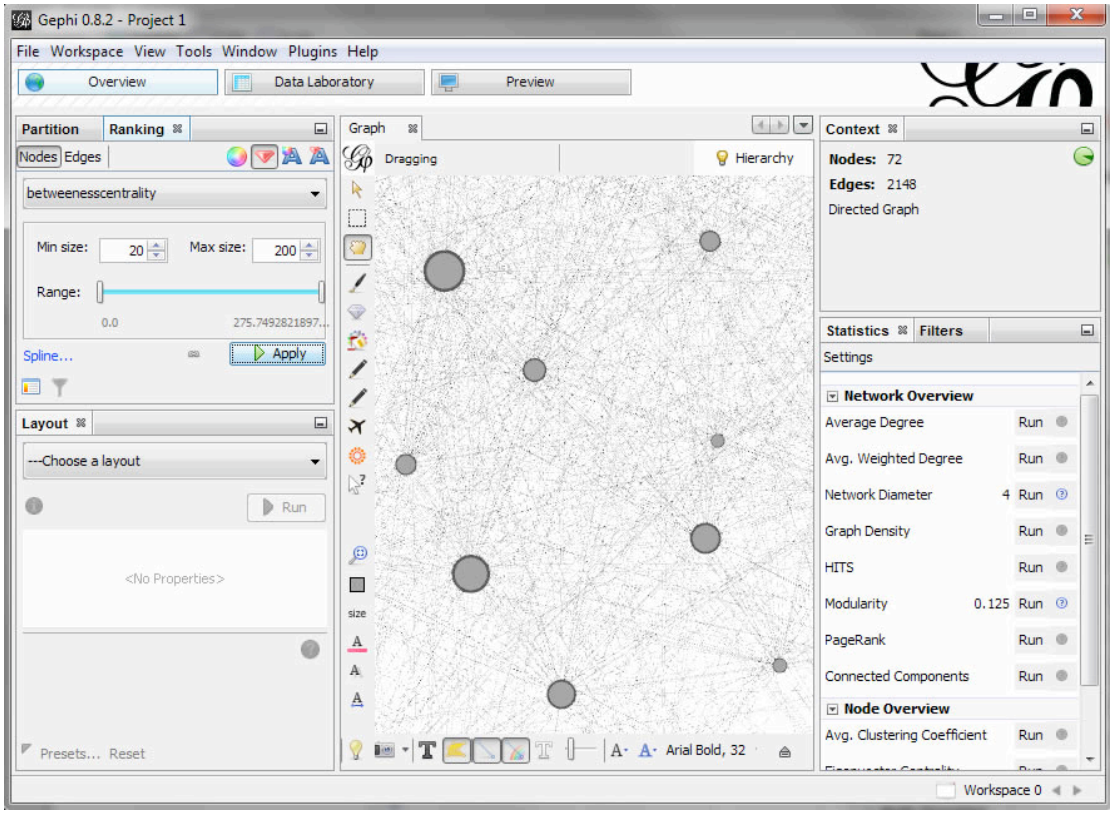
Various reports open up, but they can be safely closed without needing any attention. The value in running the tests comes from the visualisation that can now be carried out using the computed statistical values.

Finding influential individuals

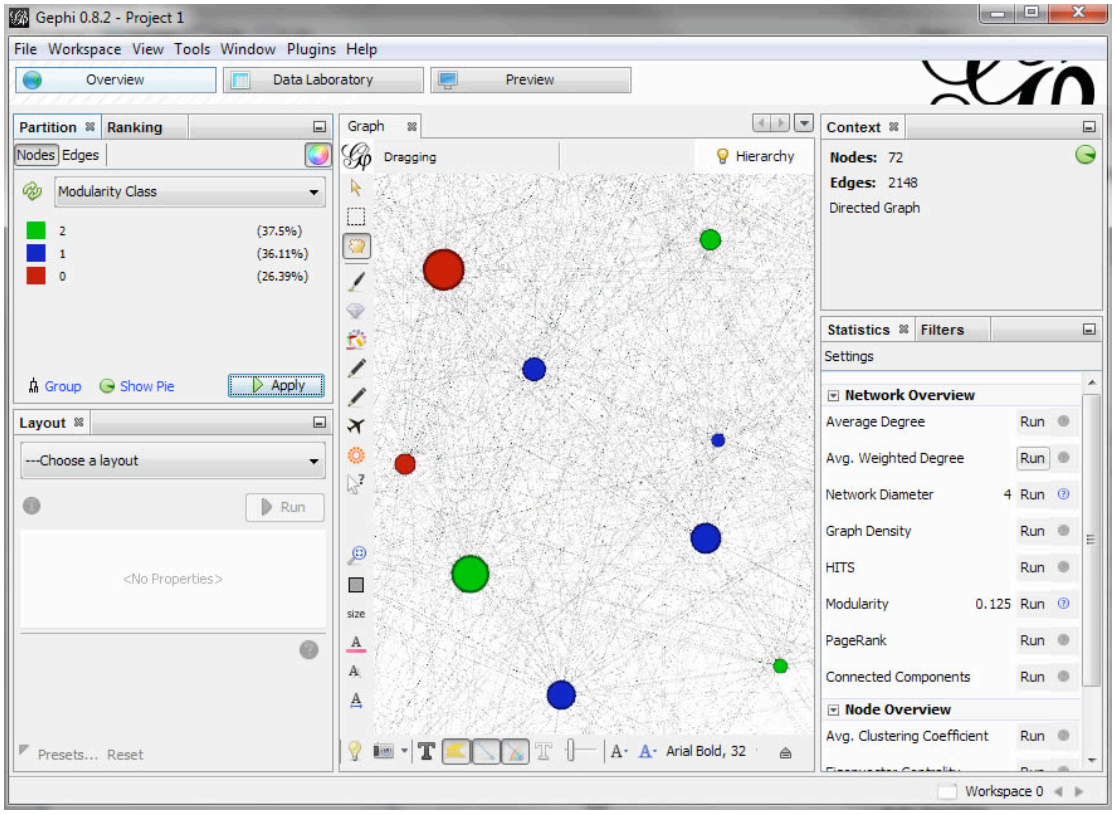
Find the 'Ranking' panel on the left hand side of the Gephi window. This panel allows the nodes within the network (each node represents an individual) to be transformed, either in colour or size. To make our most influential individuals appear as larger nodes, make sure that 'nodes' (rather than edges) and the diamond icon are selected (the diamond icon means that node size will be changed and the colour palette next to it changes the colour of the nodes). Click on the 'choose a rank parameter'. Lots of options are presented. Two important options are 'degree' (the amount of connections one node has to others, or how well connected it is) and 'betweenness centrality' (a mathematical measure of how important a node is as a step between other nodes in the network, or how many times individuals have to go through this node to get to others). This latter measure is generally thought of as a measure of influence. Select 'betweenness centrality' as the parameter and click 'apply':



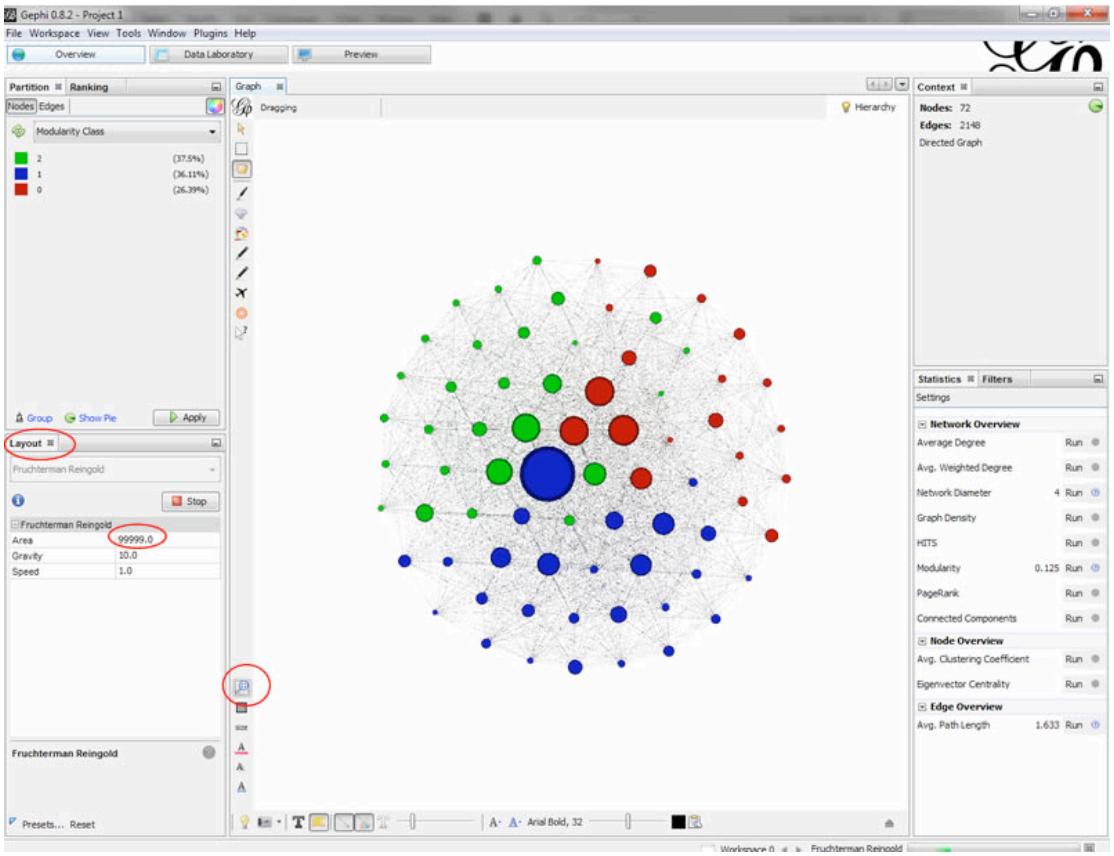
The graph should now contain nodes of different sizes – the big ones are the most influential:



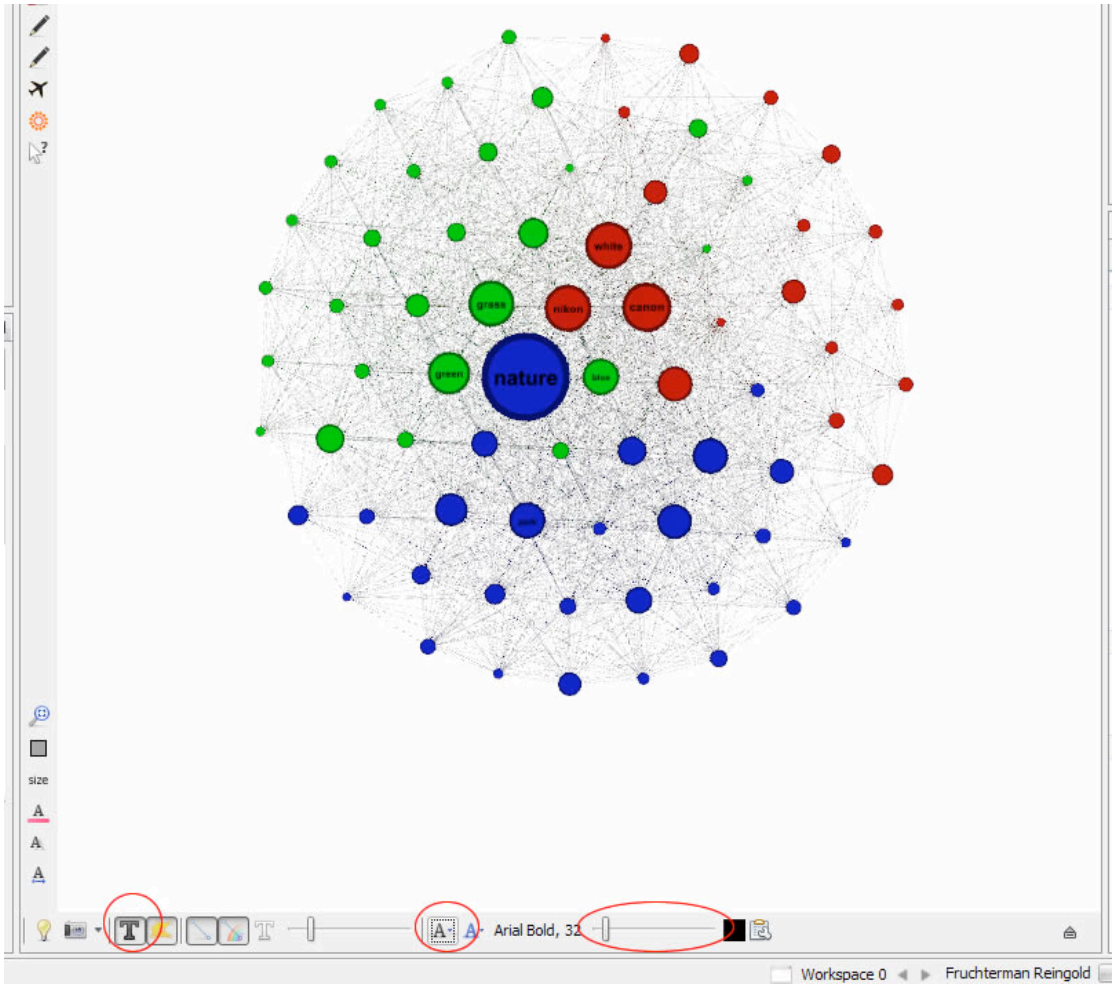
Next, find the 'Partition' panel and choose 'Modularity Class' as the parameter. Click apply. The map should now change colour, with sub-communities within the network becoming visible.



The network is now analysed and just needs a few layout changes. Find the layout panel and experiment with the different layouts. Particularly useful options include 'Force Atlas' and 'Fruchterman Reingold'. Increase or decrease the gravity or area values to make the diagram spread out more or less. There is also a handy zoom tool on the graph panel that helps to re-centre the map if it gets lost off screen:



Finally, the node labels can be made visible using the black 'T' icon at the bottom of the screen. The black 'A' icon sets the display mode – the 'node size' option is usually clearest, displaying labels only for the biggest nodes. The right hand slider controls the text size:



The most influential nodes are now visible and labelled and the communities to which they act as gateways are identified in different colours.

This kind of visualisation can often be a time-consuming process. Getting node size, label size and colour schemes right for display can be difficult. But with a bit of practice it becomes easier and provides a relatively quick way to make sense of complicated networks.

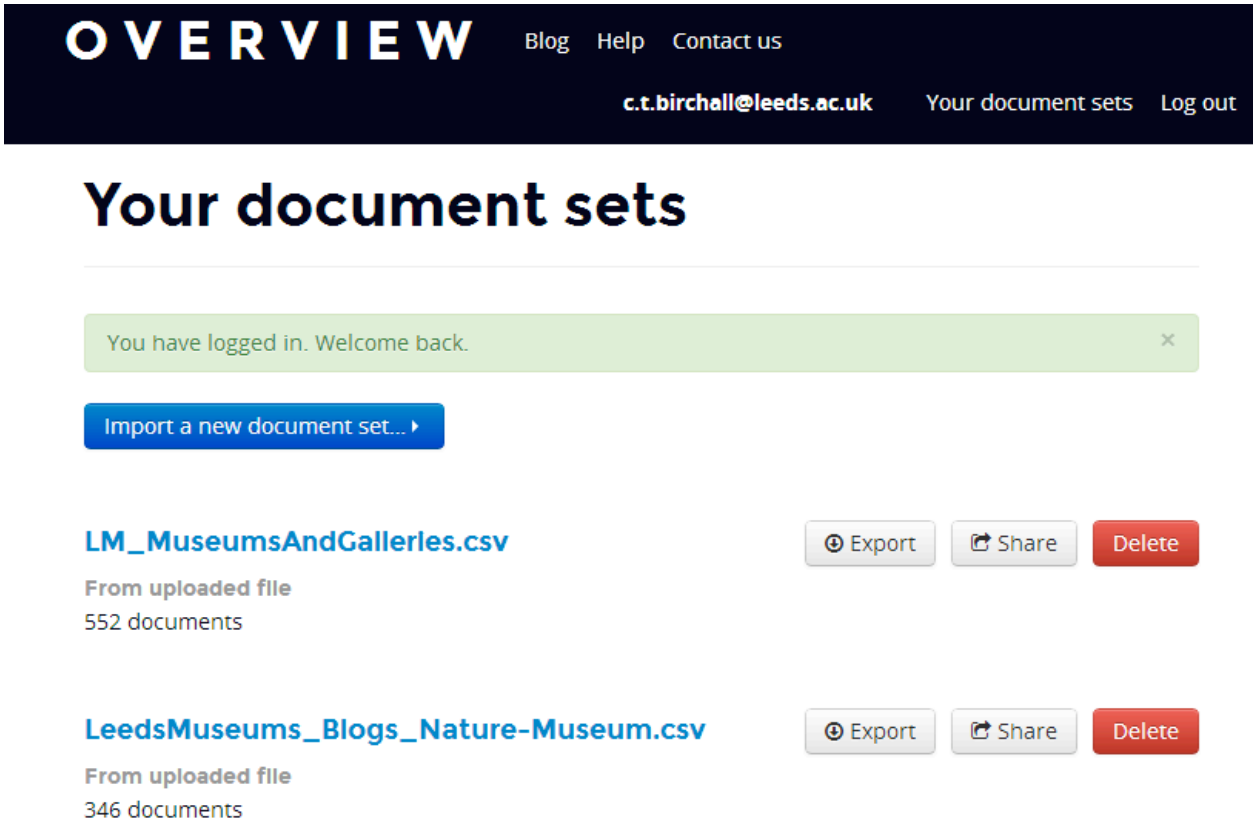
The official Gephi user guide can be found at: https://gephi.org/users/quick-start/.

2.5. Overview (https://www.overviewproject.org/)

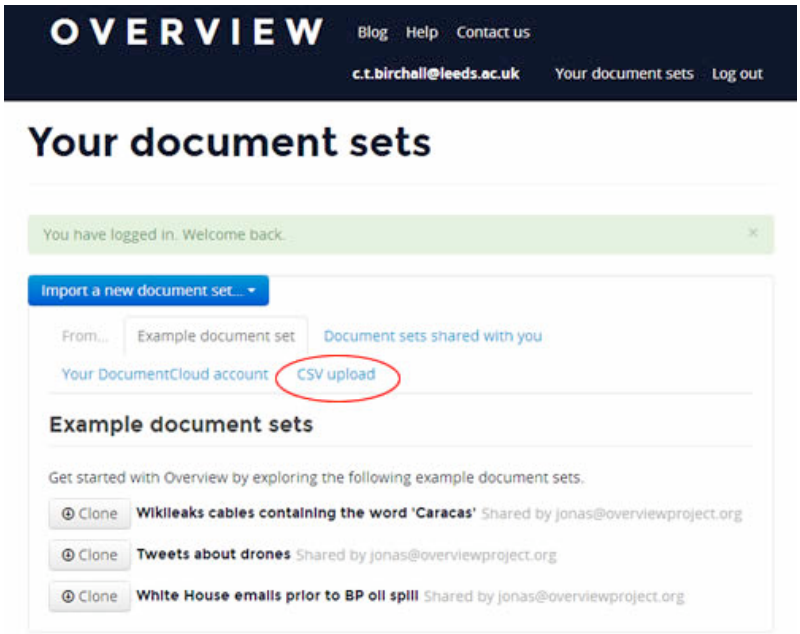
Overview is an online document-clustering tool developed by the Associated Press. The tool performs co-location text analysis and clustering, effectively grouping together documents that share words that are common amongst them, but not amongst documents in general. It allows users to identify relevant contributions and tag these.

The Overview website allows users to sign up for free, though accounts have to be activated by site administrators. Journalists and researchers are usually given access;

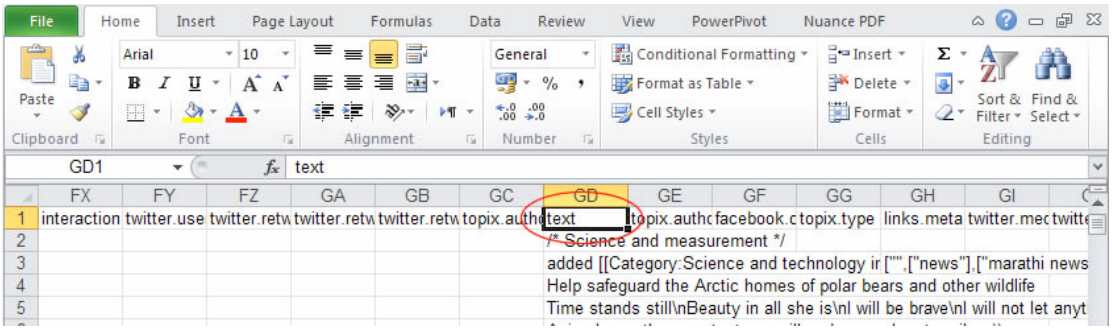
there is no published policy for commercial users, but a paid-for service to commercial users may be offered in the future. After signing up and logging in to Overview, the user is presented with a list of data sets that have been uploaded and the option to import a data set.



Clicking on 'Import new document sets' unveils a screen with options for getting data into overview. Users will normally need to upload a CSV file (such as raw data from DataSift).



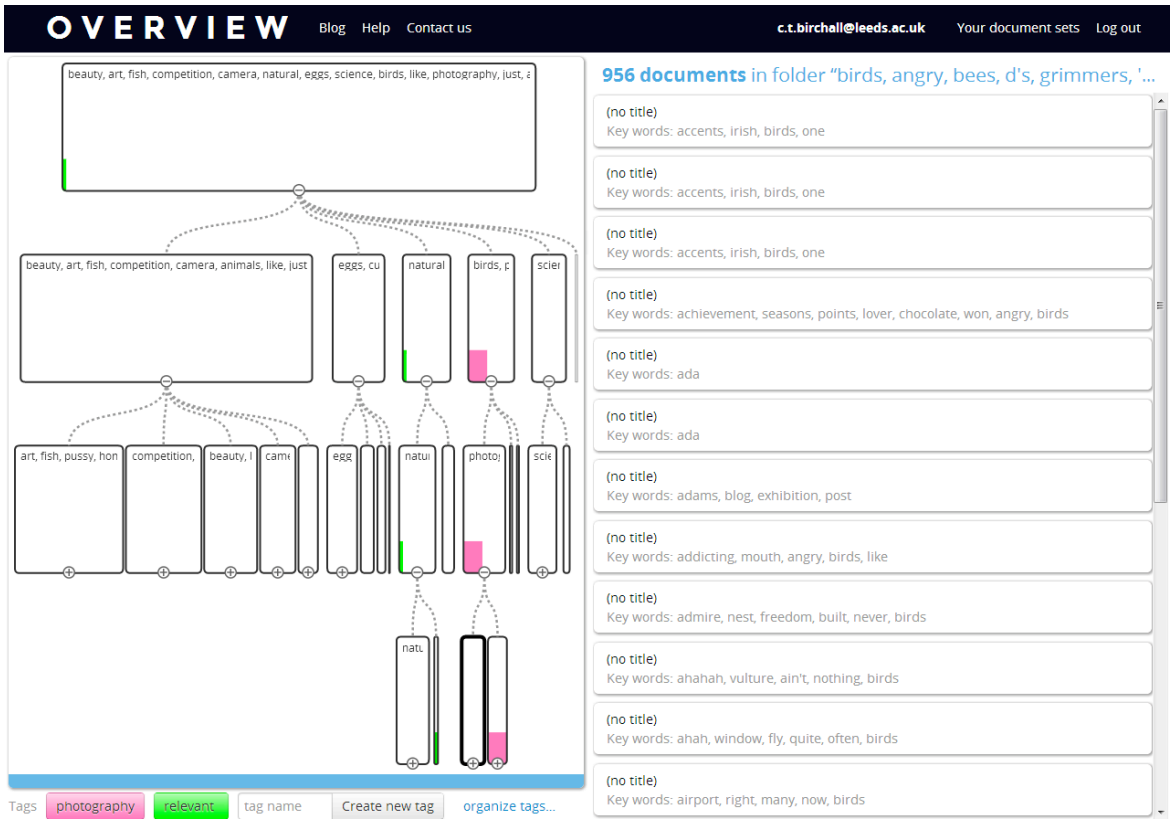
Clicking on ‘csv upload’ unveils a further dialog box with instructions detailing the exact requirements for CSV files that are to be uploaded. The most important of these are the column names in the spreadsheet, as these will almost always need to be changed. For a DataSift data set, the easiest way to prepare the file is to open it in MS Excel, locate the ‘interaction.content’ column and rename it to ‘text’:



Better still, copy this column and paste it into a new worksheet, to simplify the data set.

In either case, save the file, ensuring that it remains in csv format (not converted to an Excel document) and upload to Overview.

After uploading, Overview will perform cluster analysis of the data, grouping similar documents based upon repeated and shared content. The vertical list of contributions that existed in the CSV file is now transformed into a tree structure of grouped contributions. Each group can be expanded, by clicking on the ‘plus’ icon (+) at the bottom of the cluster, to reveal smaller clusters within. Clicking on a cluster causes the individual contributions within it to be displayed in the reading pane on the right hand side:



The clustered data set can now be tagged to identify relevant content within the data set and to add descriptive keywords that summarise the conversations within clusters. In the diagram above, clusters of content have been identified as relevant, and some clusters have been tagged with the keyword 'photography'. The sections that are not tagged can be ignored as irrelevant, letting the researcher get to relevant content quickly and easily.

Overview may be used to identify groups of relevant content within large sets of documents or contributions and these may be analysed, perhaps to identify key themes, or further filtered. For instance, the content coming from blogs could be isolated, helping to identify influential bloggers and allowing the manual categorisation of particular blogs. Alternatively, the content from UK or local sites could be tagged, filtering it from the masses of global data. This process allows human verification of relevant groups of data and human classification of them, ensuring the kind of accuracy that is not present in some automated tools. The process is time-consuming, but can add value at the investigatory stage.

3. Commercial platforms & services

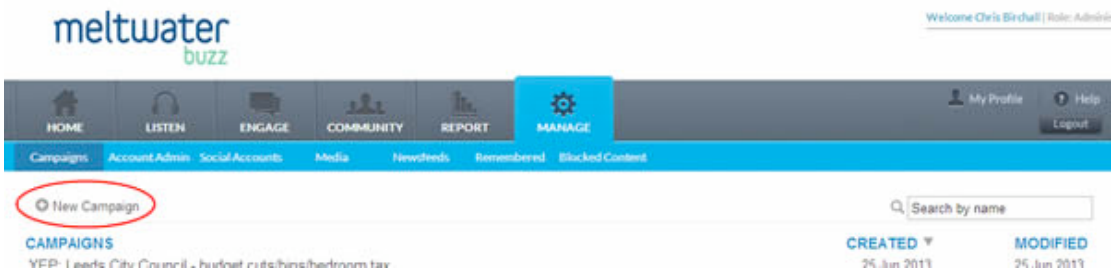
3.1. Meltwater Buzz (http://www.meltwater.com/products/meltwater-buzz-social-media-marketing-software/)

Meltwater Buzz is a commercial solution, available via paid-for subscription. The subscription cost is a one-off flat fee of £7040 excluding VAT. This allows for: up to 5 users; up to 10 search campaigns (unlimited keywords); and training, support, set-up and consultancy. There is some flexibility in tailoring costs; contact Sheena Norquay (sheena.norquay@meltwater.com) to discuss this. Meltwater Buzz does have usage limits - it can be 'throttled' if a search is too wide and returns too much data.

This document briefly covers usage of Meltwater Buzz. Further guidance and support is available from the vendor for paying customers.

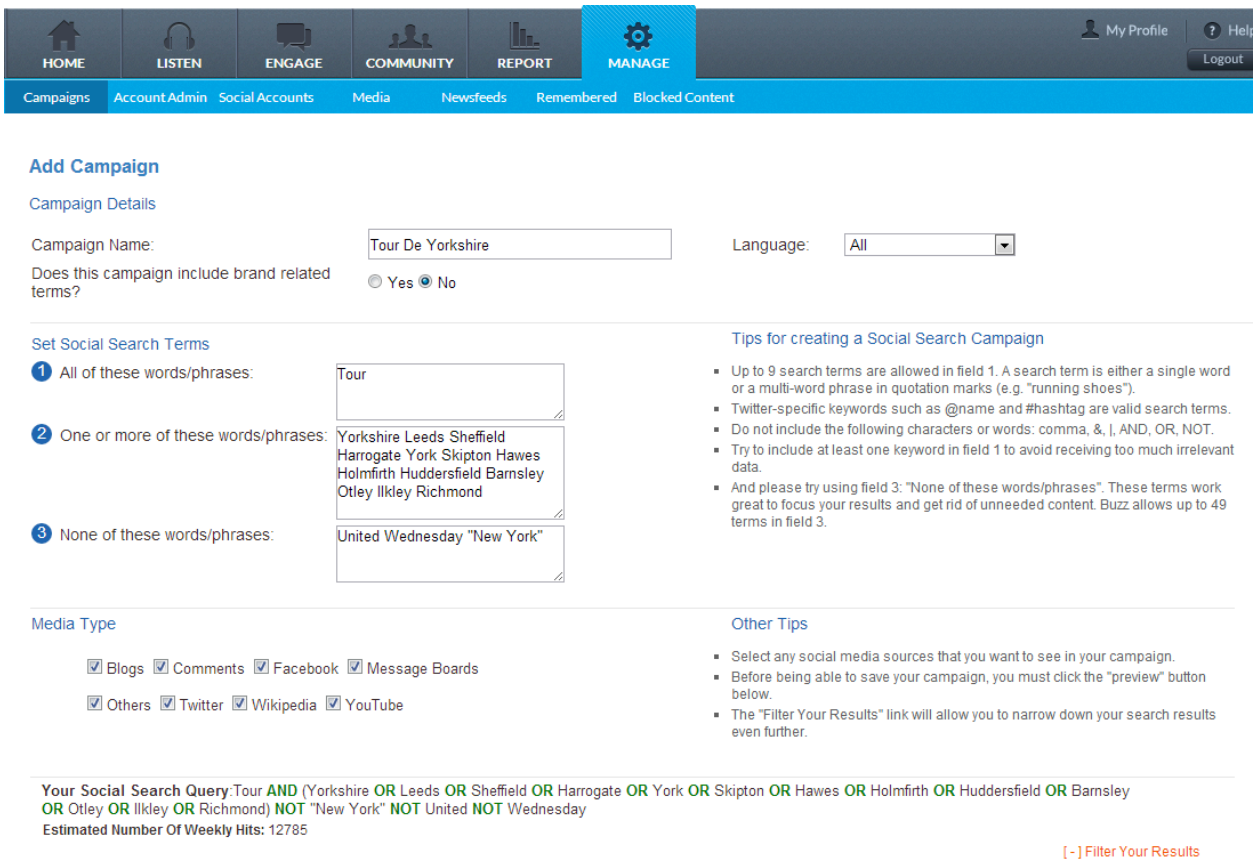
Meltwater Buzz is the social media analysis tool that complements its parent product, the more traditional media-monitoring tool, Meltwater. Because of this, Meltwater Buzz only harvests content that it defines as 'social'. It does not look at general online information, such as newspaper articles or other web pages, as this is done by the original Meltwater product. Rather, it looks at what people are saying about certain keywords. The product accesses a comprehensive range of social data sources including Twitter, Facebook, blogs, comments boards, message boards, Wikipedia, Youtube and 'Others'. While it does not harvest complete web pages (unless they are blog posts), as it regards these as non-social media, Meltwater Buzz does harvest individual comments posted on web pages (such as news articles), which is particularly important for those interested in what local people are saying about specific issues. Meltwater Buzz allows the user to tailor the data sources accessed through the ability to block content so that particular data sources do not appear in your results (which is useful for getting rid of irrelevant data). It is not possible for the general user to target searches at specific data sources (such as particular local forums) but it may be possible for the tool administrators to do this. Meltwater is relatively easy to use, though the Boolean search system requires training for those who have never used it before. This search interface provides a flexible and powerful way to filter results.

After logging in the user can create a search (known as a 'campaign' within Meltwater Buzz) by navigating to the 'Manage' tab and clicking 'new campaign':



The 'new campaign' provides various places to enter criteria that will shape the data returned by the search. The steps involved take the user through the whole process, first defining the larger scope of a search and then adding fine details in the form of a filter to remove any data that is likely to be irrelevant to the topic under investigation.

The first step is to set the main search terms. As in DataSift, the user can specify words that must be included, may be included or must not be included; there is a separate box for each of these options. The search below looks for all contributions containing the word 'Tour' in combination with a variety of Yorkshire place names, but excluding any contribution that contains the terms 'United' 'Wednesday' or 'New York' (removing a lot of irrelevant football and US-related data). Note the specific guidance to the right of the boxes:



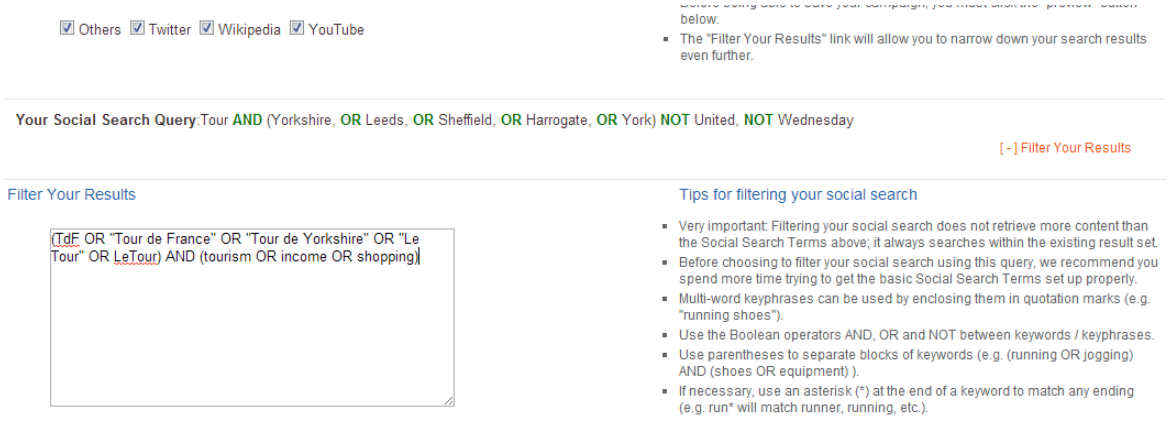
Note that underneath the boxes the search query is written out in Boolean language (AND, OR, NOT, etc.).

The next step is to choose the data sources to be targetted – note the list available under 'Media Type' in the ilustration above. By default all data sources are checked.

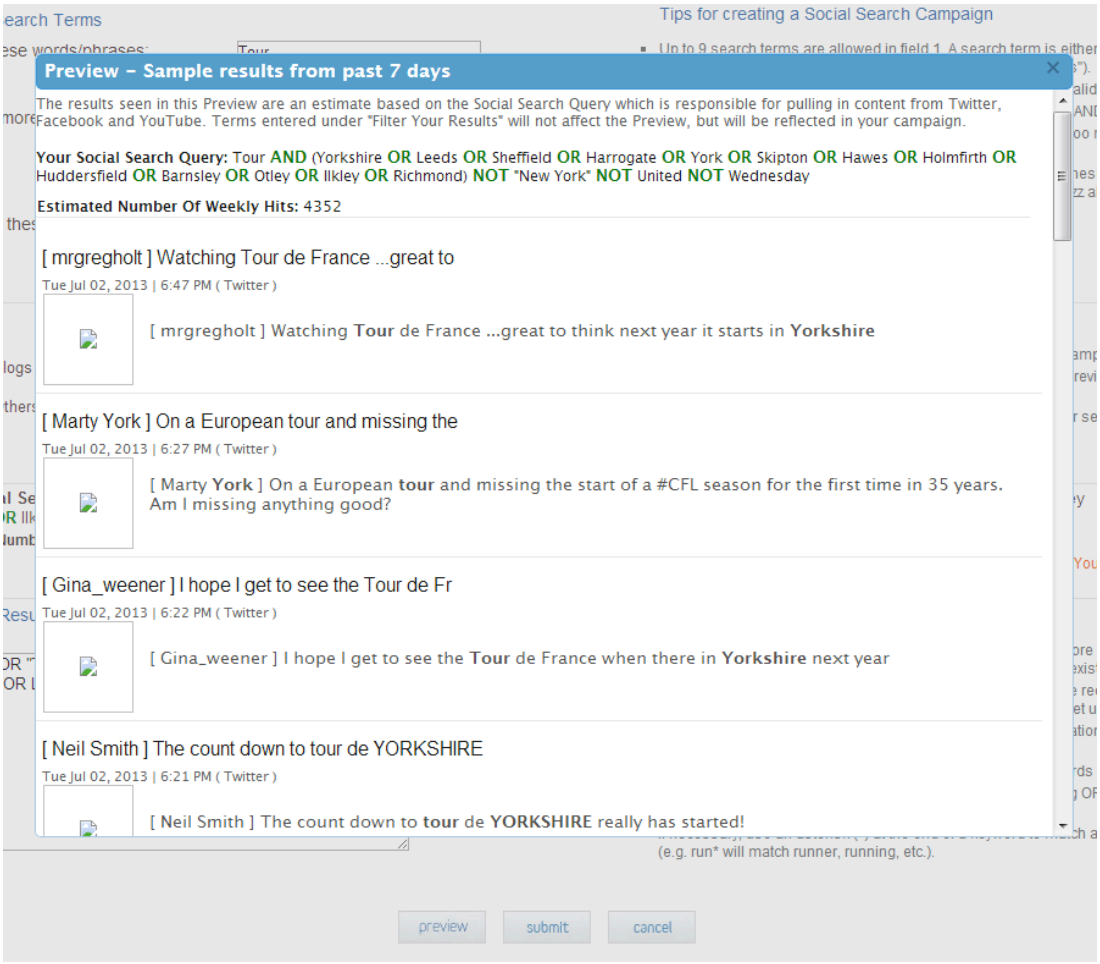
The third step is to add a specific filter to minimise the number of results gained, fine tuning the data set to be really relevant to the topic to be investigated. A filter can be added by clicking the 'Filter your results' link under the search definition boxes, and typing in a Boolean query in the box that appears. In the diagram, below the filter requires contributions to contain some more specific Tour de France terms as well as some terms related to tourism and shopping. Thus the result set can be reduced from

general talk about the Tour in Yorkshire to talk specifically about the Tour de France in Yorkshire and shopping, tourism and income. Note the specific syntax of the filter, including parentheses and double quotes. Boolean search queries like this are very logical, but also fussy about syntax.

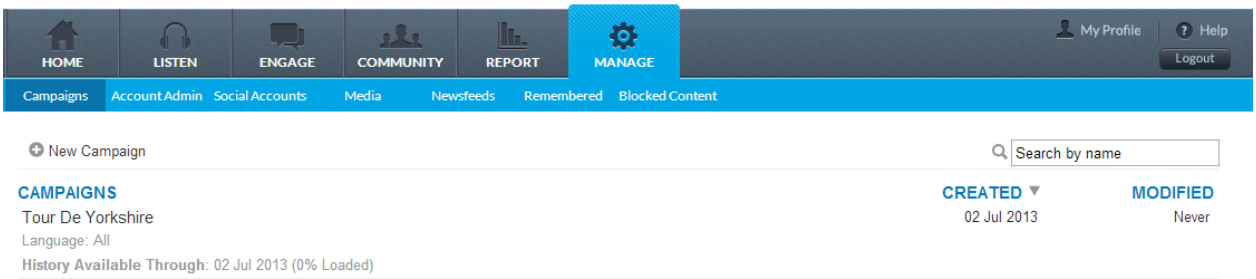
Note how the two types of terms – Tour de France and Tourism/Shopping are each encapsulated in parentheses and seperated by the keyword AND. This means that the contributions must contain words from both sets to appear in the results. The words within each group are seperated by the OR keyword; this means that contributions only need to contain at least one of the terms from each group to appear in the results.



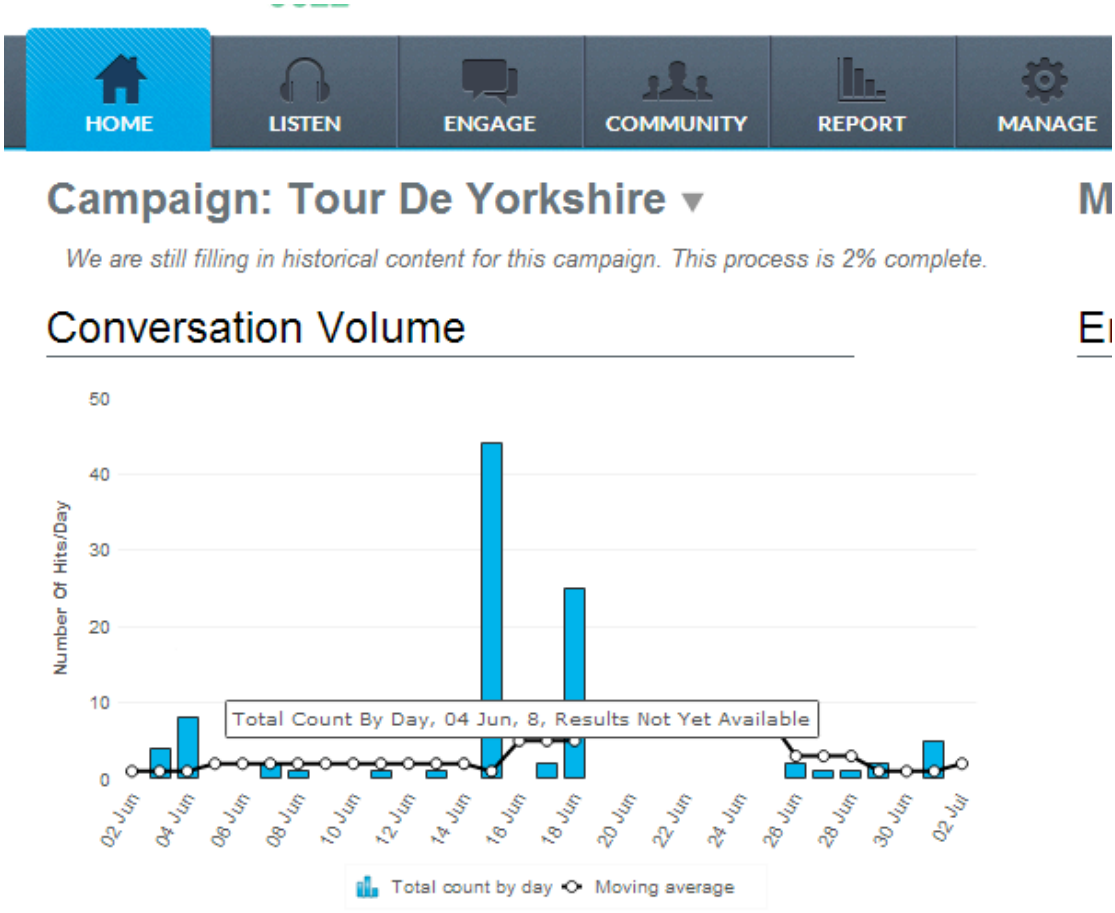
Once the search terms have been defined the search can be previewed. A pop up box shows the number of expected results along with a sample of contributions. This preview does not utilise the filter defined above, just the search terms from the three boxes at the top. The filter will reduce the results that appear, but the tool will actually harvest the number shown in the preview. The preview can be closed and the user can either add more search terms to narrow or widen the search, or click submit to save the search and begin the buiding process.



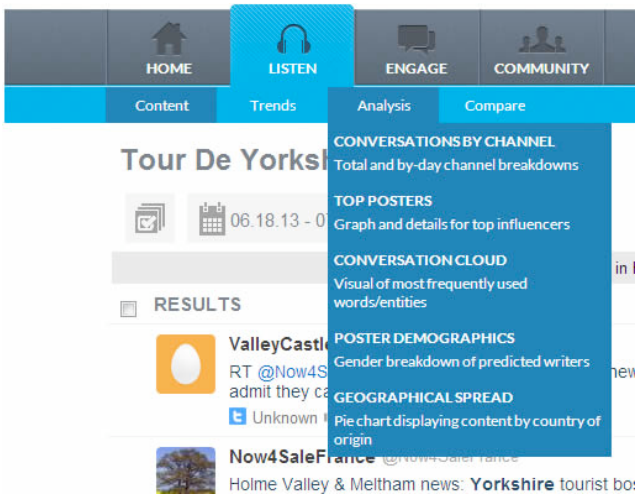
After clicking submit, Meltwater Buzz sets about building the query and harvesting results. The user is redirected to the 'Manage' tab and the new search appears with a progress report showing how much data has been collected. The process often takes a few hours.



Once a search has been defined, various different visualisations are created to help with analysis of the data. On the 'Home' tab, graphs illustrate headline figures, such as a chart showing how the results were created over time, allowing a picture to be built up of how online conversation volume has changed, identifying peaks and troughs of activity which may be related to particular events, announcements and press releases. At each point on these graphs, the user can view the contributions at a particular date to see the content of the conversations occurring.

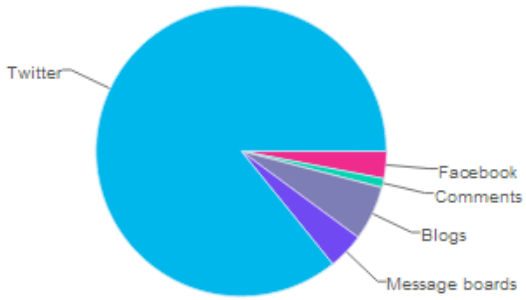


In the analysis tab, Meltwater Buzz also shows graphics that identify many other statistics and trends, such as the source of the contributions (showing the proportion of search results returned from different channels, such as Twitter, Facebook, blogs, etc), a list of 'top posters' and a conversation cloud showing the most commonly used word.



Within the first of these analyses, the contribution source, the tool does not show statistics about the websites within each channel category, so while forums may be identified as an important source, the user would have to look at the individual contributions to work out which particular forums were most important. Both of these two visualisations need to be taken in the context of the specific Meltwater Buzz data

sample. Some data sources may be represented more strongly in the results due to the tool having better access to them. For instance, Twitter has well indexed data and is therefore accessible, while blogs and forums are often much more difficult to access. It is important to work closely with Meltwater Buzz to ensure that the most important data sources are included in the sample, such as local forums that might not make it into the default sample. Preliminary research to identify these data sources is therefore vital.



Meltwater Buzz presents a range of other information such as gender and location of contributors within particular search results. These are based upon very incomplete data sets – typically over 90% of contributions do not have gender attached and over 80% have no geographic data attached. Meltwater Buzz will also attempt to determine the sentiment of contributions (another very inexact science) but does seem to have success in summarising conversations by extracting key terms from them and presenting them as a content cloud.

Meltwater Buzz works by harvesting content (that is, contributions from users, based on keyword matches). It does not target individuals and then analyse social networks from that starting point. Looking at influencers in Meltwater Buzz is the equivalent of looking at the social media accounts in a DataSift result set and then analysing the interactions between them, in contrast to the analysis of key accounts through NodeXL. However, looking at results in list view, you can order by date, views (based on message impressions) and rank (based upon Social Rank and Impression data, a bespoke 'black box' algorithm created for Meltwater Buzz). You can also view a table of top posters and you can view the profile of users across platforms, including their accounts on Twitter, Facebook and other platforms, their number of followers, their conversations (contributions). In the 'Community' tab you can view and edit the profiles of the user that you have added to your database, adding custom notes if necessary.

3.2. Brandwatch (http://www.brandwatch.com/)

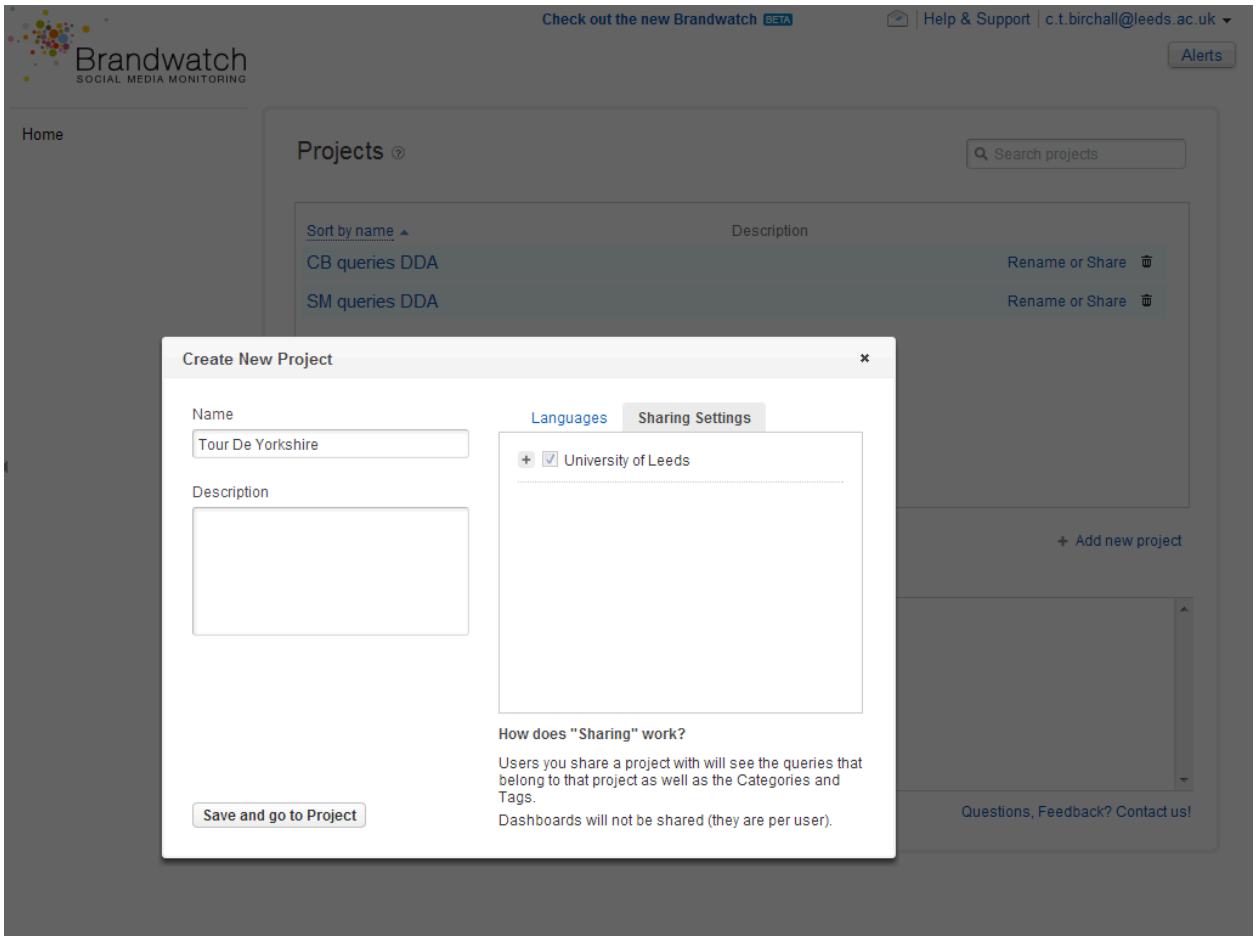
Brandwatch is a standalone commercial package for monitoring the online footprint of brands, campaigns or topics. Harvesting all types of online content, from news articles to tweets, Brandwatch can offer a rich picture of what is being contributed online in relation to any topic being investigated. There are three different base packages offered by Brandwatch, which allow access to differing quantities of data and to a varying range of support services:

- Brandwatch/Pro (Social analysis for small and medium brands or topics), £500p/month

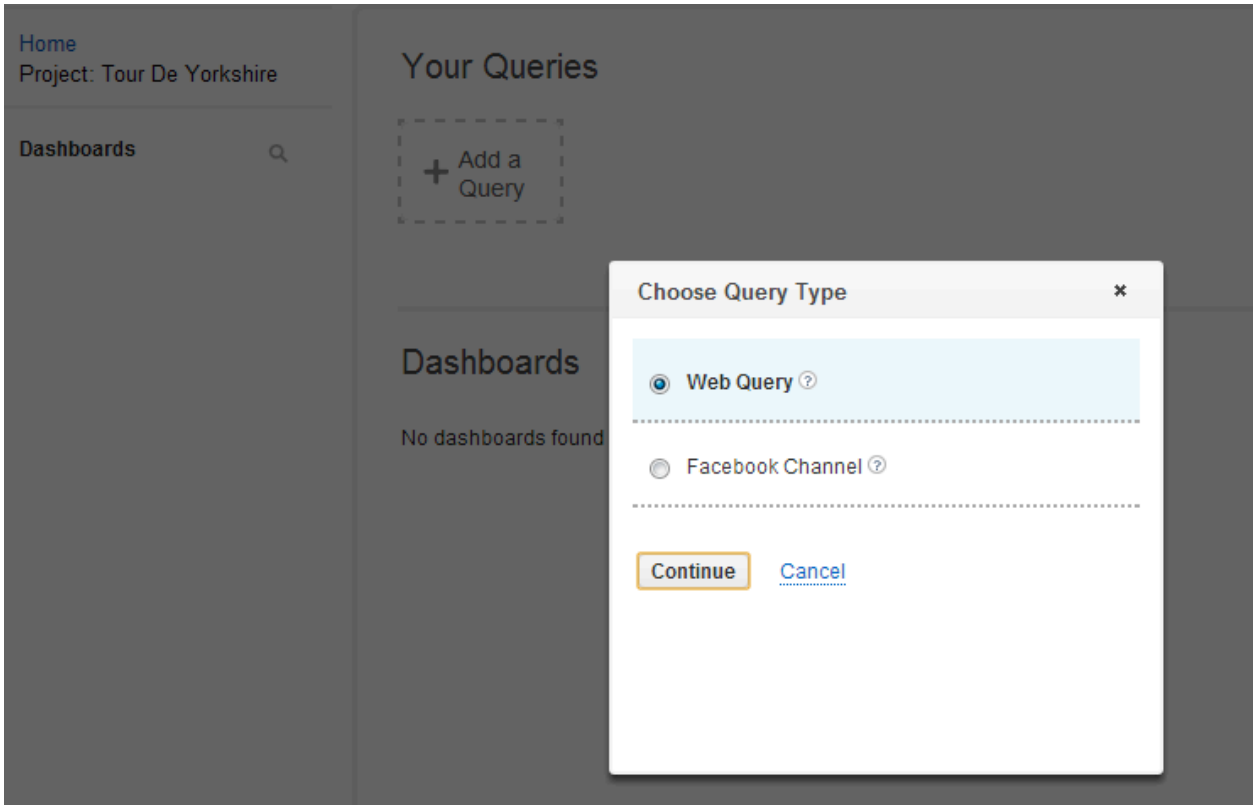
- Enterprise/M (Advanced analysis for big brands and large agencies), £2000 p/month
- Enterprise/Q (Advanced analysis for high-volume queries), £2000 p/month.

More information can be found here: http://www.brandwatch.com/why-brandwatch/pricing/.

To create a new search, first create a project by clicking on the 'new project' link on the homepage after logging in. Projects hold multiple searches (or queries as they are known in Brandwatch). Users from an organisation can share projects and they can even be set to use different languages:

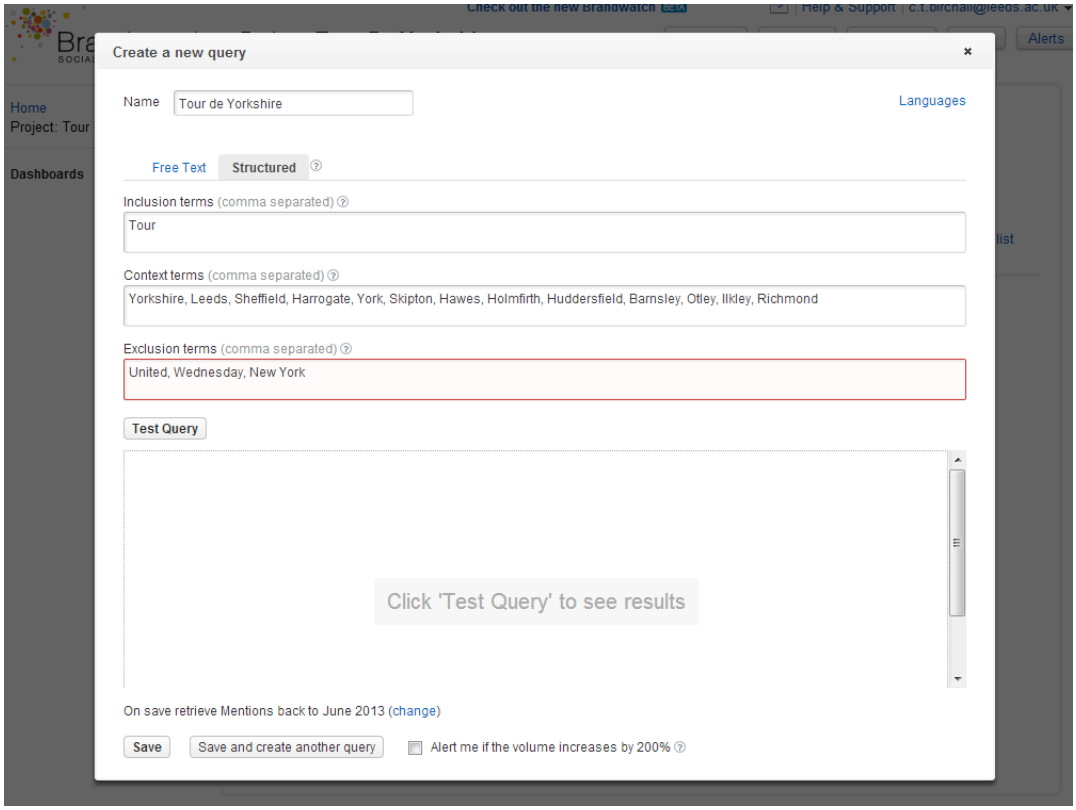


Choose a web query type in order to harvest a range of data:

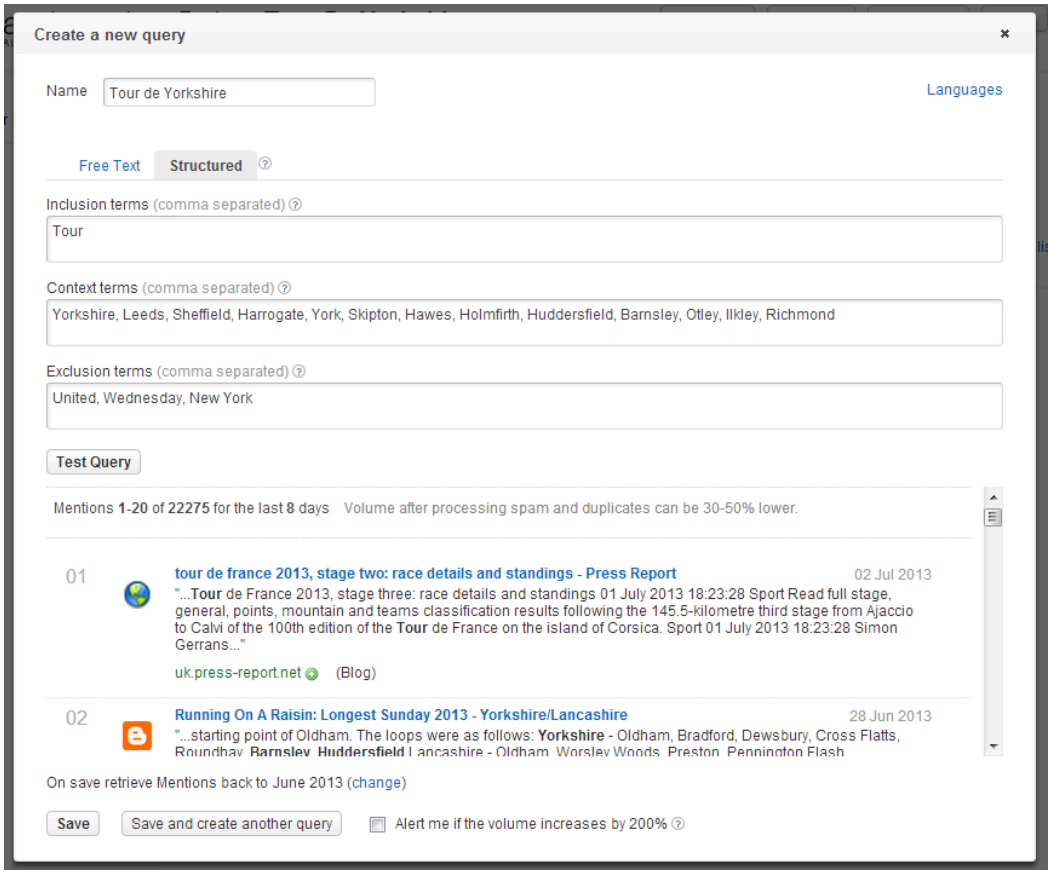


Then add the search terms. Like Meltwater Buzz, there are three boxes to enter ketwords into. Labelled 'inclusion terms', 'context terms' and 'exclusion terms' these are analogous to the 'must contain all', 'must contain at least one' and 'must not contain' boxes. In the search defined in the illustration below, the same search terms are used as in the Meltwater Buzz example (above). Note the similar-but-slightly-different interface and syntax.

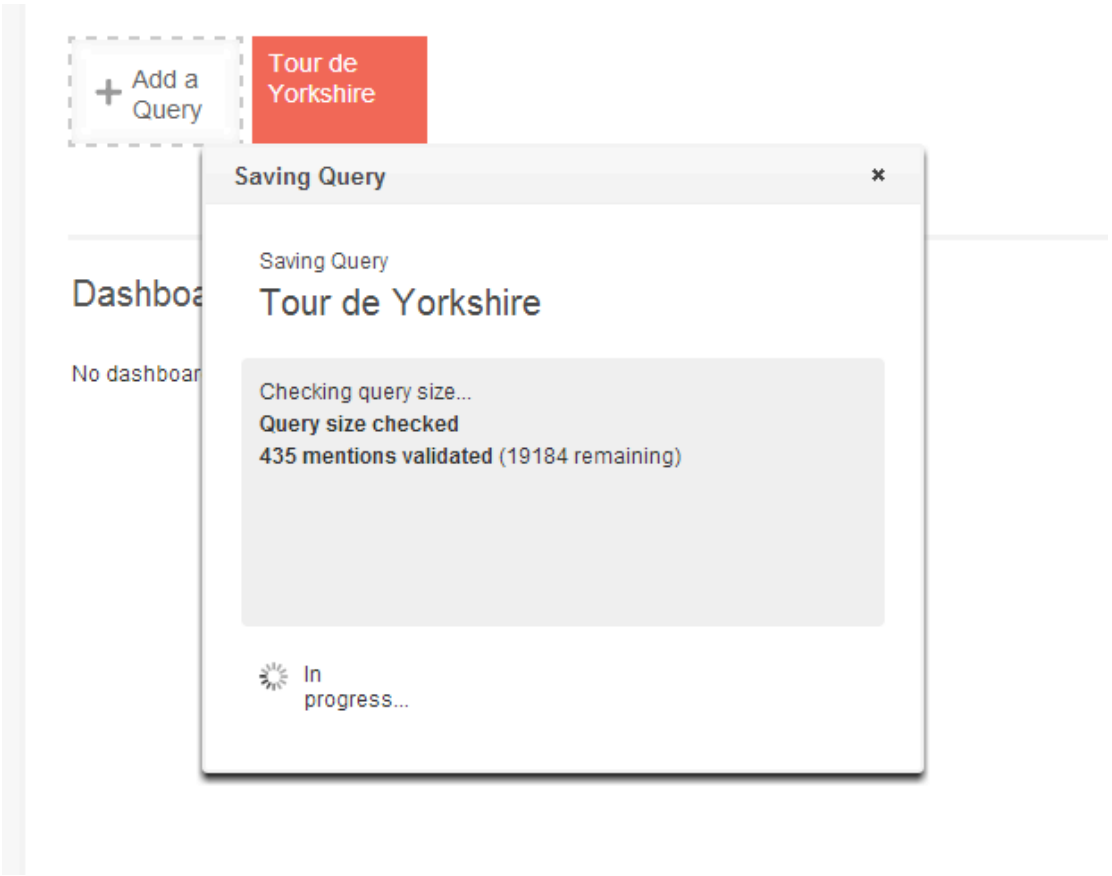
Brandwatch also offers a free text option where Boolean search strings can be typed in manually, in the same way as in other products (including Meltwater Buzz) ensuring that knowledge of Boolean syntax is a transferable skill across all of these products.



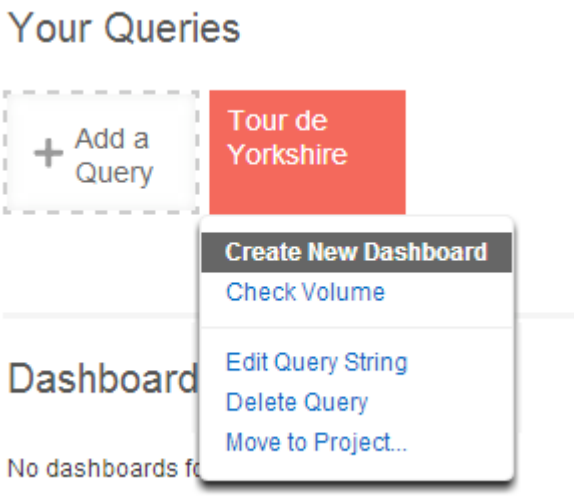
After defining the search, click the 'Test Query' button and Brandwatch will retrieve a sample of contributions as well as an estimation of the amount of data that the search will find:



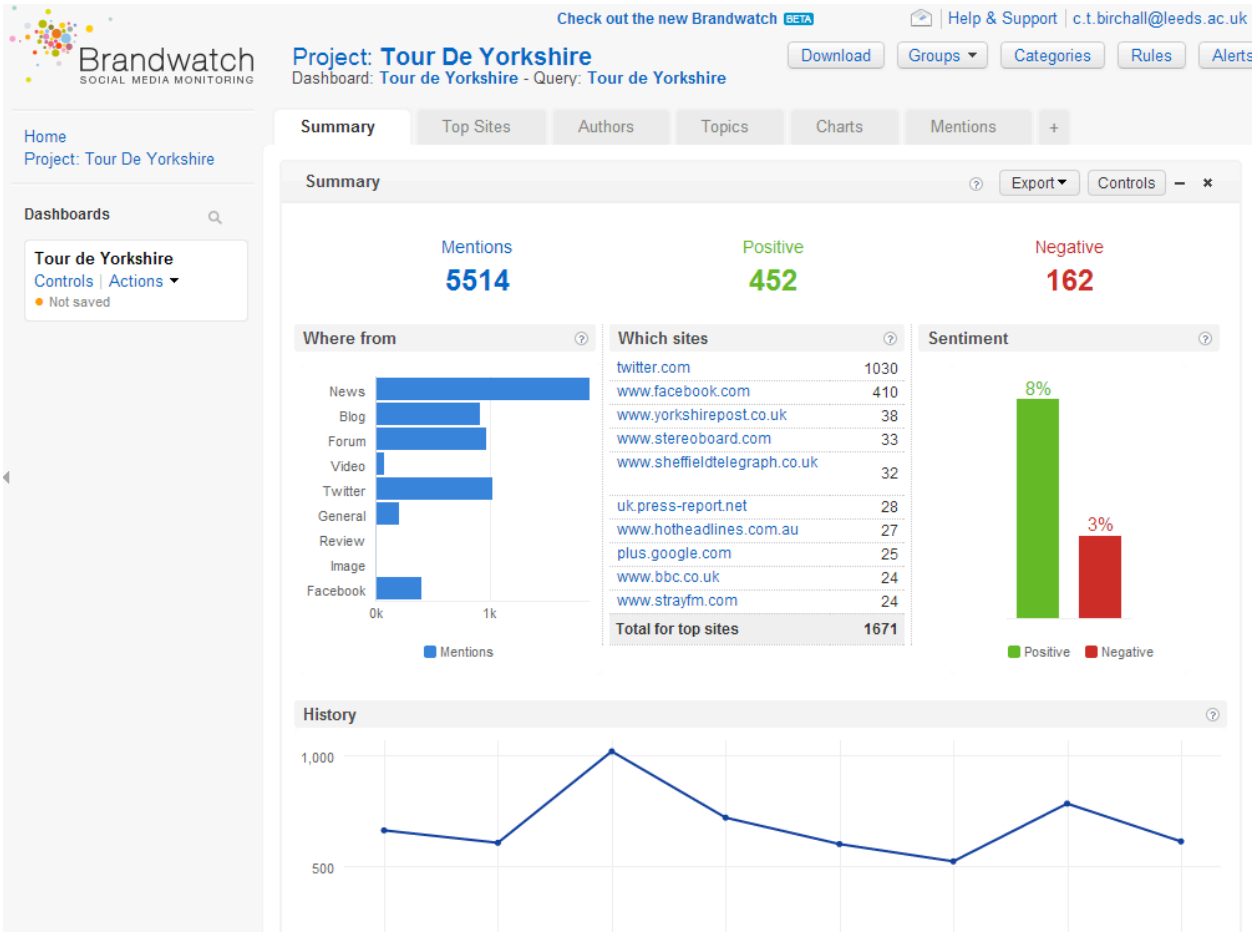
When you are happy with the search, click the 'Save' button and the process of building and executing the search will begin:



After a few moments, the search is ready for viewing. Clicking on the new query brings up a list of options, allowing the user to edit the query, check its volume (in case it is too big) or create a new dashboard:



This latter option is the most exciting – dashboards are the screens in Brandwatch that hold all of the visualisations and statistics:

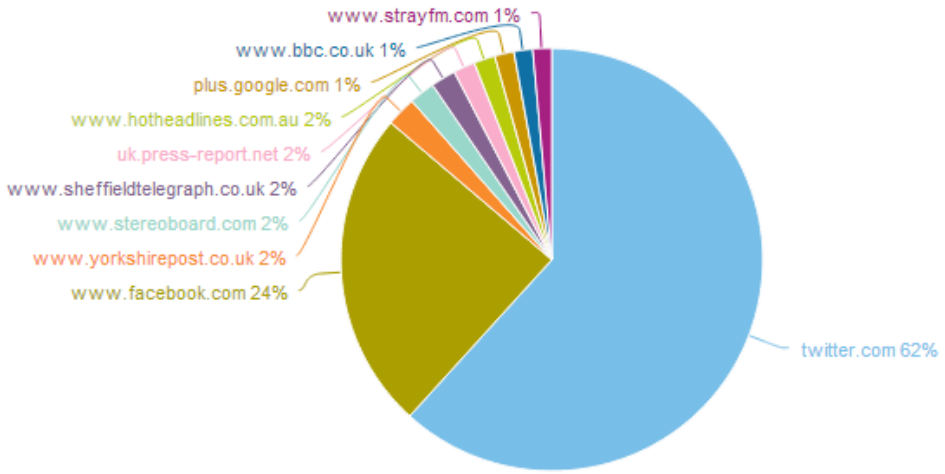


Brandwatch offers a similar range of visualisations to those offered by Meltwater Buzz and the constraints are the same (for example, results are shaped by access to data sources).

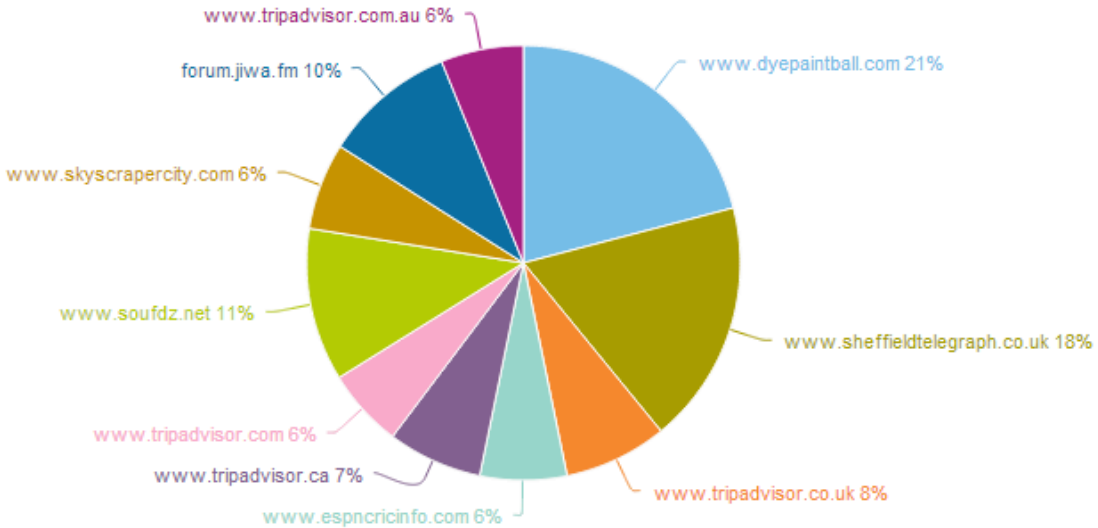
However, there are two key differences between the two services. Brandwatch harvests all web content, not just social (Meltwater Buzz just does the social, leaving the rest to its parent product, Meltwater), and because of this it does not separate out comments from articles on web pages, instead classing the whole thing as one contribution. Meltwater Buzz does not harvest the actual article (unless it is a blog post) but harvests each individual comment as a contribution.

Brandwatch identifies specific platforms within channels (see the illustrations below for 'top sites' and 'top forums' for the 'Tour de Yorkshire' search) and this allows it to identify specific local platforms such as *LeedsForum.com*, *SheffieldForum.com* and *skyscrapercity.com* as important sites of discussion. These sites do not appear in headline results in Meltwater Buzz, just the generic channel of 'forums' (interestingly, looking at the raw results, Meltwater Buzz only seems to find data from Leeds forum).

Top sites identified by Brandwatch for one search:



Top Forums identified by Brandwatch for one search:



3.3. Other commercial platforms & services

For the purposes of our project we only reviewed the two commercial services discussed above: Meltwater Buzz and Brandwatch. These are both well-known with the social media monitoring industries. However, there are numerous similar services available, offering some of the same features, and some distinct features too. The costs of these services are often not disclosed in online marketing materials. Some of these are listed below.

Two global and well-known companies are:

- Radian6, a well-known, global product, recently acquired by Sales Force Marketing Cloud (http://www.salesforcemarketingcloud.com/). Basic,

Professional, Corporate and Entreprise packages are available at undisclosed rates.

- Sysomos (http://www.sysomos.com/), another global company offering a range of products, including HeartBeat monitoring and MAP (Media Analysis Platform) analytics.

Other companies whose marketing materials indicate that they have experience of working with public sector organizations include:

- comScore Social Essentials (http://www.comscore.com/Products/Audience_Analytics/Social_Essentials)
- Matraxis (http://www.matraxis.co.uk/)
- SEOPositive (http://www.seo-positive.co.uk/).

Other companies include, but are not limited to:

- Synthesio (http://synthesio.com/corporate/en)
- BrandsEye (http://www.brandseye.com/)
- BuzzCapture (http://www.buzzcapture.com/)
- Social360 (http://social360monitoring.com/)
- Lithium (http://www.lithium.com/)
- JellyFish (http://www.jellyfish.co.uk/).