Chapter 13
Advances in Crowdsourcing: Surveys, Social Media and Geospatial Analysis: Towards a Big Data Toolkit

Steven Gray, Richard Milton and Andrew Hudson-Smith

Abstract The collection, mining and analysis of social media are arguably one of the core examples of "big data" sets for the social sciences. The dynamic nature of the media makes it a new and emerging base for the analysis of human behaviour and brings new opportunities to understand groups, movements and society. Analysing the results of billions of conversations has already revolutionised marketing and advertising. However, these datasets, by their very nature, are complex, time-consuming and computationally difficult to analyse. We put in place a series of examples to utilise such datasets with a view of exploring non-complex workflows via the use of new toolkits, linking into data collection via the crowd and opening up systems for analysis.

Keywords Crowdsourcing · Geospatial · Toolkit · MYSQL · Mapping · Geographic · SurveyMapper · MapTube · Tweet-o-Meter

13.1 Introduction

A key method to process large datasets is to outsource basic analysis to human volunteers and, in some cases, large groups, to help process this data by manually identifying patterns, features or interesting events within the datasets. This process has been called "Crowdsourcing", first coined in the article "The Rise of Crowdsourcing" (Howe 2006b).

S. Gray (✉) · R. Milton · A. Hudson-Smith
The Bartlett Centre for Advanced Spatial Analysis,
University College London, London, UK
e-mail: steven.gray@ucl.ac.uk

R. Milton
e-mail: richard.milton@ucl.ac.uk

A. Hudson-Smith
e-mail: a.hudson-smith@ucl.ac.uk

F.J. Garrigos-Simon et al. (eds.), *Advances in Crowdsourcing*,
DOI 10.1007/978-3-319-18341-1_13

"Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers" (Howe 2006a).

Crowdsourcing in the realm of a read/write Web (known as Web 2.0) is the process of collecting individual actions that can be aggregated together to create a collective result (Hudson-Smith et al. 2009). Projects such as Galaxy Zoo (Sloan Digital Sky Survey 2010) have had notable success with this process for astronomical datasets from sensors collecting data from sunspots to radar images from distant galaxies. Members of the public have successfully identified candidate galaxies in different solar systems through these highly specialised, custom-built systems by splitting up huge datasets into small pieces, a process that would take traditional forms of research a number of years to analyse. These crowdsourced systems have proven to be extremely important as the amount of data collected has increased due to the rise of real-time sensors and live streams of situational data that make up the deluge of data available today (Demarest 2011).

However, many crowdsourced projects suffer from a lack of users or simply not enough members of the public willing to help through other means, such as lack of exposure or a poorly designed interface. Good exposure of a project is vital to getting users to contribute data and, more importantly, getting users to join the community. Crowdsourced projects need a crowd—so that new data are being generated, processed and outcomes obtained.

Services such as Amazon's Mechanical Turk (Amazon 2011) and those such as Solar Stormwatch (Zooniverse 2009) use volunteers who are either interested in the subject matter or offer some sort of financial reward to entice users to use these services. Users in the Mechanical Turk ecosystem offer a price per action, a Human Intelligence Task or HIT, for a worker, to carry out an action such as categorising an image or translating a passage from one language to another (Ross et al. 2011). If the worker is qualified to carry out the HIT, then after completion, their associated Amazon account will be credited by an amount from $0.02 for a simple task up to $20 for a more complicated task. This is one way to get users to carry out tasks, but puts up barriers to research, as many projects cannot afford to provide monetary compensation to users.

Researchers have explored these rich datasets of social information available through the open APIs as a by-product of the rise in popularity of social media services. Twitter officially released an interface to the data for the service 6 months after their initial launch due to the number of third-party developers creating applications from Twitter data by automatically "scraping" pages of data rather than using a machine readable API (Stone 2006).

In this chapter, we will document the various experiments and toolkits created at the Centre for Advanced Spatial Analysis (CASA) for crowdsourcing data and look to the future of crowdsourcing using modern computing techniques to automate collection. The first such example we explore is SurveyMapper.

13.2 SurveyMapper

Collecting data on a global scale require outreach beyond the traditional methods. Traditionally, services are built from the ground up to support the collection of data for specific needs rather than a general need for many users. In late 2009, CASA was approached by the BBC to carry out a real-time survey for broadcast exploring the anti-social behaviour around the BBC Look East region (Norfolk, Suffolk, Essex, Cambridgeshire, Northamptonshire, Bedfordshire, Hertfordshire and Northern Buckinghamshire, UK) using the popular MapTube platform, allowing the visualisation of both survey and geographical data. At that time, few, if any, technologies existed for collecting real-time crowdsourced data in this fashion and so a bespoke application built on top of MapTube was created that allowed a single survey to be carried out with the results updated every 30 min. This survey resulted in 6902 responses overall with an average of 1340 results generated after broadcast over 3 evenings. MapTube was developed as part of the Generative e-Social Science project (GeNESIS). It provides a dynamic API (called MapTubeD) to render data into raster tiles and the ability to create and clear tiles on request. As a user provides a response to a survey, the system updates a database of responses and alerts the MapTubeD system to fetch a new datafile. When the map is next requested on the page load, the new raster files are downloaded from MapTubeD and placed on top of the survey map. These requests appear transparent, and the complexity is abstracted away from the user who only sees an updated map of responses in real time. It became apparent after the BBC survey that there was a need for this type of generalised platform and this bespoke project became the genesis of SurveyMapper.

The application, defined as "SurveyMapper", is a platform, which allows users to set up their own surveys and collect geographical locations from participants along with their views on the survey (Fig. 13.1). It was specifically created to tap into the crowd and visualise the results not only geographically but also in near real time. To set up a survey, users are asked to provide some metadata about the survey such as a title, description, keywords for searching, and start and end date along with the questions they want to ask participants. What makes SurveyMapper different to other platforms, such as Survey Monkey, is that a survey is tied to geographical boundary areas and visualised as a choropleth map of responses to a particular question. There are seven possible geographical types a user can select to visualise data: countries, US zip codes, UK postcodes, EU countries, US states, UK counties, London boroughs or London electoral wards. In addition to these geographical areas, there are 2 point-based visualisations that users can select to provide locations to responses: a latitude/longitude marker that is placed on the map and a heat map visualisation. Tens of thousands of inputs can be collected quickly, providing a near real-time view of research questions, instead of the 30-min delay of the original MapTube system. The data and subsequent visualisation are updated immediately. Data can be exported later for more rigorous analysis or integration with existing datasets.

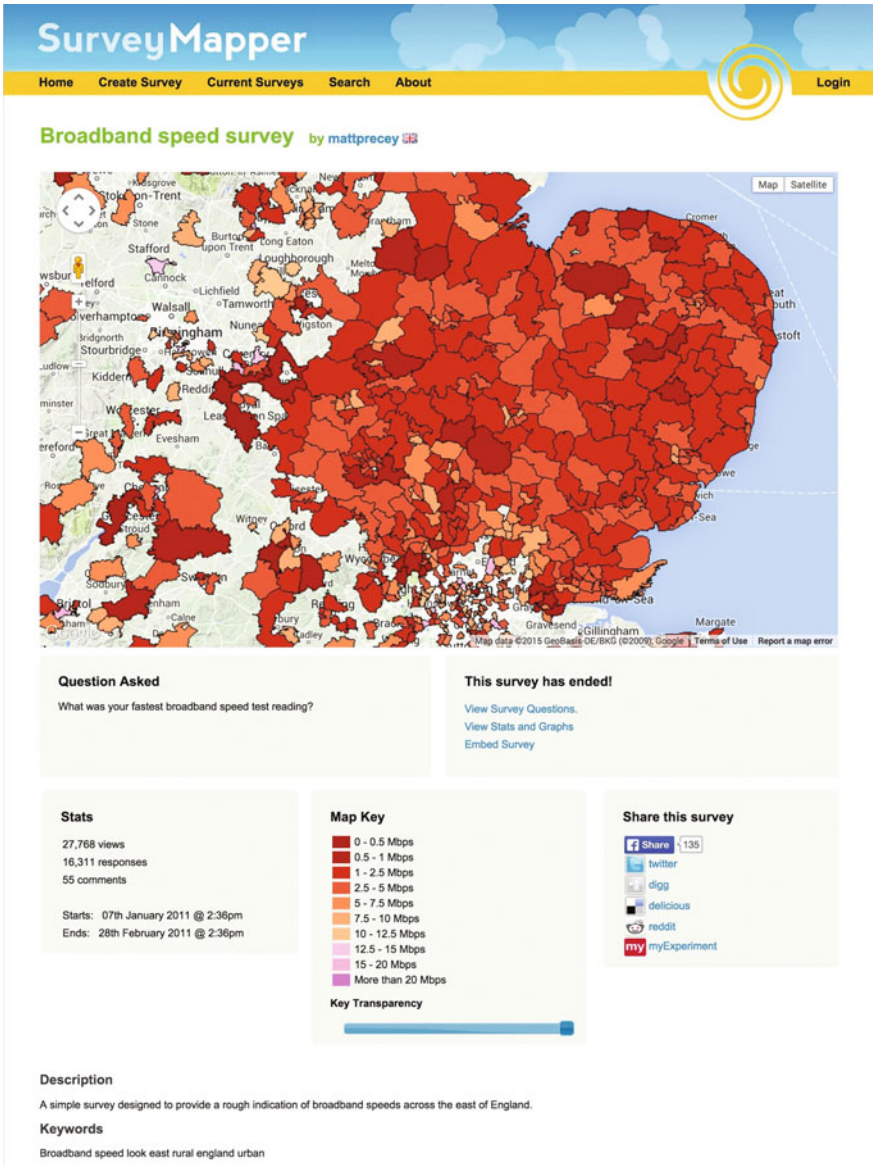

Fig. 13.1 SurveyMapper Survey-Broadband Speed Test January 2011. *Source* own

SurveyMapper can also be extended to provide custom visualisations for surveys that need specific and specialised information capture. The Greater London Authority approached the SurveyMapper team to help carry out a safety survey of parks around London. They wanted to ask users of the park, "How satisfied or dissatisfied are you with the quality of your local parks and green spaces?" through

12 questions that probed citizens' views on their local park. The issue with this type of survey is that citizens do not necessarily know the name of the local park, which is held by the GLA. By adding a London park dataset, sourced from green spaces data on OpenStreetMap to the MapTube rendering system, we were able to customise the experience of SurveyMapper to allow users to select their local park by dragging a pin around the Greater London area and dropping the pin on a parkland area. When the pin was dropped on the map, the name of the park was populated automatically and submitted to the database with the remainder of responses. At the end of the survey, the GLA downloaded the individual and aggregated data of responses and were able to integrate the data with their own internal datasets.

SurveyMapper was built with 2 specific purposes: to provide social scientists with a series of online tools to collect and visualise data in near real time allowing the creation of "mood maps" linked to a backend geographic information system, and to remove the traditional academic look to scientific software services. SurveyMapper achieves the latter by "skinning" the application and focuses on usability by removing the complexity and the science of the software from the user. Throughout the Web application, users are presented with a fun and bright design (Fig. 13.2), which incorporates the mascot "Roger the Giraffe", named after Roger

Fig. 13.2 SurveyMapper home page—non-academic branding, Roger the Giraffe mascot. *Source* own

Tomlinson who is commonly acknowledged as the "father of GIS" (Greiner 2007). This design decision was taken early on in the process of creating SurveyMapper to give the system a look and feel more akin of a Web 2.0 project rather than a scientific academic project.

As users set up a survey, they are encouraged to share the survey on social media platforms as we have found the most successful surveys on the platform use social media to reach participants. The system provides short URLs and custom links too as well as links that allow users to embed surveys on personal Websites and share with social networks such as Facebook and Twitter. The system also allows users to comment about individual surveys and to promote discussion on the platform about results of surveys. Collecting this data allows research into and users of the application to flag up interesting results within the dataset. To date, more than 500 surveys have been created with a combined total of 152,310 responses and 210,000 views in the first 5 years of SurveyMapper being available to the public.

During October to December 2008, CASA ran a survey to gauge the public response to the proposed congestion charge in Greater Manchester. This asked the question:

If you have an interest in how the £2.7 billion plan to reshape Manchester's transport system will affect your neighbourhood then here's your chance to add what you think to an interactive map of the region. This online collaboration between BBC Manchester and experts at the University of Manchester will give a unique picture of how well the proposals are going down across the northwest. Simply select one of the options listed below, enter your postcode and click on the submit button.

MapTube will include your answer in the next new map.

"If a congestion charge was brought in would you:

Drive and pay the charge?
Drive at different times?
Use public transport/motorbike/bicycle?
Work or shop elsewhere?
Not Affected?"

This resulted in a total of 15,902 responses during the three-month period with the final data are represented in Fig. 13.3.

Due to the politically sensitive nature of the issue and the fact that there were a number of pressure groups who opposed the congestion charge, the 45 % of people responding "Work or shop elsewhere" is perhaps not that surprising. This also highlighted the major problem with crowdsourced surveys, because certain members of the general public were deliberately submitting large numbers of responses in an attempt to manipulate the survey. This was something that was expected and is relatively simple to filter out the data.

By selecting and counting duplicate IP addresses and then mapping the data by postcode sector, the map (Fig. 13.4) clearly shows "OL8" and "M31" with many more responses for "Drive and pay the charge" than any other postcode. Our only explanation for this is that somebody was trying to make the predominantly blue

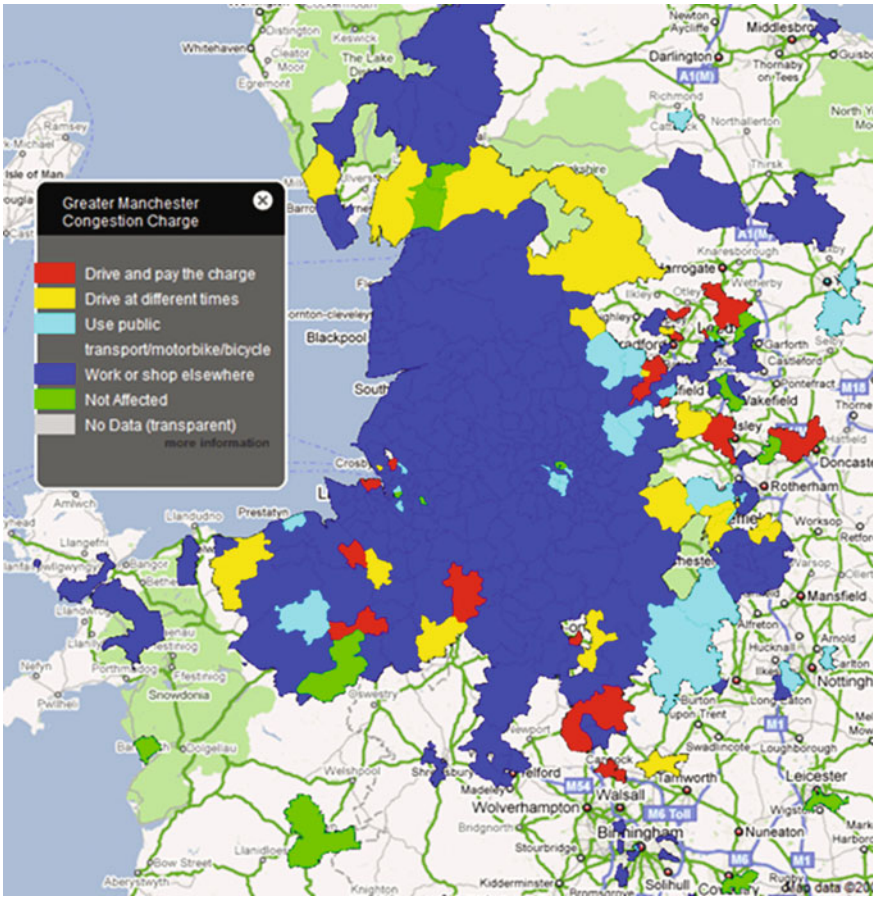

Fig. 13.3 Greater Manchester Congestion Charge Survey—live map can be viewed at the following link: http://www.maptube.org/map.aspx?mapid=239. *Source* own

map turn red for his or her own postcode. Geographic density of responses is an important issue with georeferenced surveys, as the non-uniform nature of the responses needs to be taken into account when analysing the data.

SurveyMapper has preventions built into the core of the system to prevent users, as well as identifying, from gaming the system and adding extra responses to a survey. By default, each survey only allows a single IP address to vote on a survey. If a user is logged into the system, their response is logged and the interface will prevent the user from viewing the "Enter Response" page. As SurveyMapper also allows users to vote from embed surveys, third-party apps that implement the SurveyMapper API as well as users who are not logged in, extra precautions have been implemented to detect irregularities in patterns of response. A database table of all responses, which serves as a time series master list, is checked nightly for patterns of burst activity in the remote chance that a user has defeated the response

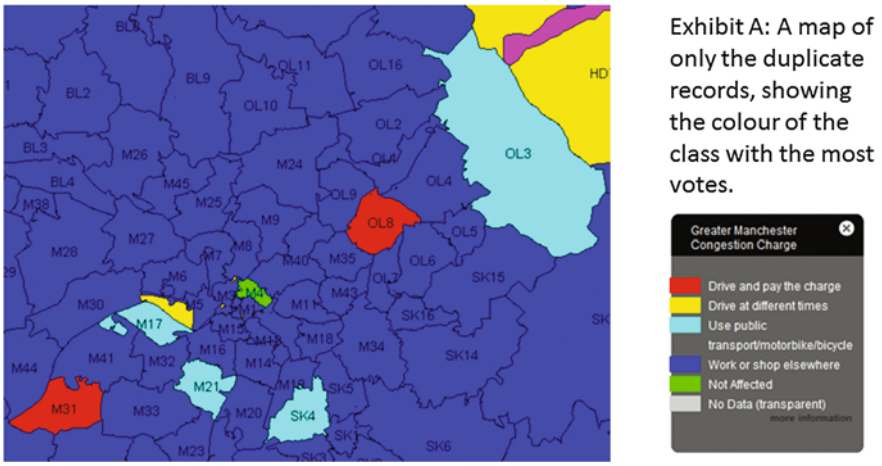

Fig. 13.4 Selecting only the data for IP addresses which were used more than once reveals some interesting results. Further analysis shows hundreds of responses entered for "M31" and "OL8" from two addresses. *Source* own

lockout event. Surveys that are currently live maintain a count of individual responses for each individual question as well as a value that is visualised on the map. As soon as a survey ends, a transient table is created and the results are rebuilt from the master list removing any responses that seem suspicious. This table is then archived to save space on the server and to speed up responses to the survey for future visits. This method of validation ensures that the results returned to the user, via the direct data download, have been checked for accurate survey results that have been provided to the survey creator.

13.3 Tweet-o-Meter

While exploring use cases of the SurveyMapper system, the team identified that to promote usage of the platform outside the Website, an API would be needed to submit responses to surveys from third-party applications and that allowing users to quickly submit responses without visiting the main SurveyMapper Web application. Twitter, a social media platform, that enables users to send messages of 140 characters or less proved to be the best solution to allow users to quickly respond to a survey. It was envisioned that a user would learn of a survey from posts within their social network and be able to submit a response by either retweeting a post, the process of sharing a tweet with your own network, or creating a post with the survey id and the response value.

This method of posting real time using tweets was popularised during the winter of 2010 when the UK experienced an unusually prolonged spell of severe cold weather that resulted in heavy snowfall (Met Office 2011). Twitter users started to organically

share regional snow reports throughout the UK in real time by rating the current snowfall on a scale from 0 to 10 and including the first half of their postcode within the tweet. As this trend started to spread, various Websites appeared mapping the real-time results on top of online maps. One such Website was UK Snow Map (Marsh 2014) created by Ben Marsh, a freelance Web developer. At the same time, the team at UCL CASA started to experiment with collecting the UK Snow tweets from Twitter and mapping them using the choropleth mapping technique used for SurveyMapper.

Figure 13.5 shows the total number of tweets collected during January 2010 when some of the heaviest snowfall was taking place. Of all the tweets collected, 36.4 % included a postcode ("Steady heavy snow in Chesham HP5 #uksnow"), while 28.8 % included a postcode and an amount ("#uksnow SG10 3/10"). This exposes a basic problem with this type of crowdsourcing, because not everybody has the same opinion about the amount of snow. One person's 5/10 might be another person's 8/10, depending on where they live in the country and how much snow they are used to. Using the synoptic data from the Meteorological Office, the reported depth of snow in centimetres was compared with tweets in the same postcode, leading to the conclusion that there was no measurable correlation between the two (Fig. 13.6). However, when looking at the general question of whether there was any snow in the area, the crowdsourced data compared more favourably with the official Met Office snow amounts.

Twitter exposes the data stored within the system via two authenticated APIs: the Streaming and the Search API. The Streaming API allows developers to make a connection with the API and receive updates automatically as they happen. A persistent connection is made to the API endpoint, and data are fed to the third-party program when a user sends a publically available tweet. The Search API requires a third-party program to request the data from the API endpoint repeatedly over a period

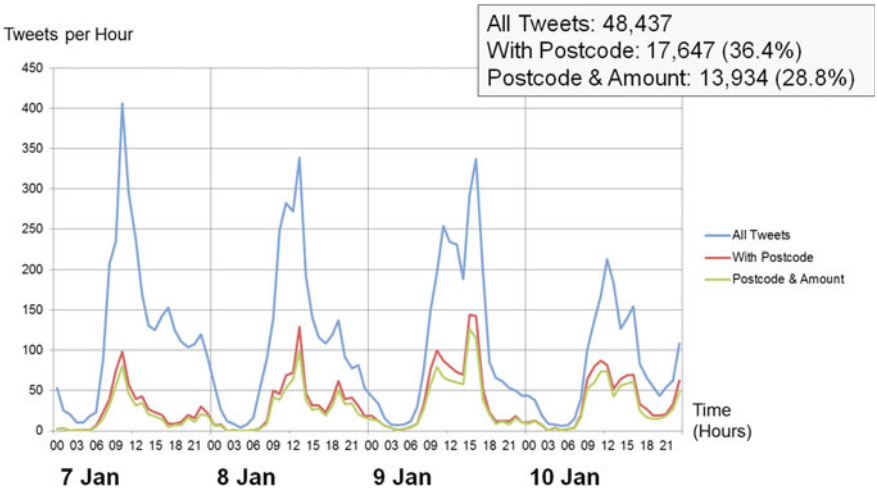

Fig. 13.5 Tweets collected from UK geoboundary during 2010 snowfall. *Source* own

Fig. 13.6 Data for 6 January 2010 12Z, the map on the *left* shows postcode sectors where there is snow on the ground using the Meteorological Office data. The map on the *right* shows the #uksnow counts from Twitter. *Source* own

of time. As the data are requested, the API returns the last available tweets (up to a maximum of 100 tweets) depending on the search query passed to the API. By making multiple requests over a period of time and tracking the last message collected, the third-party application can build a subset of tweets over time. The API serves tweet data in 3 different capacities: the Firehose, the Garden Hose and the Spritzer. The FireHose represents an unfiltered, raw feed of all tweets globally for the search query provided. The Garden Hose is a 10 % sample of all tweets queried, and the spritzer represents a 2 % statistical sample of data. Both the Firehose and Garden Hose require prior agreements with Twitter for access, which is assessed on a case-by-case basis; therefore, the spritzer feed is commonly used by third-party applications. Unlike the Streaming API, the Search API is rate limited by the number of requests within a period of time. At the time of development (2010 version 1 of the API), this limit was undocumented, but through experimentation was calculated at around 350 requests per hour. As per version 1.1 of the Twitter API, this rate limit has been extended to 180 requests per 15-min period—720 requests an hour.

During the experimentation phase with SurveyMapper and the Twitter API, we started archiving and storing the tweet and relevant metadata returned from the API using several machines inside the laboratory. Each machine ran several processes to collect data from the Search API asynchronously and aggregated the data within a central data repository for analysis at a later date. Automating the data collection between multiple distributed machines in this fashion, we were able to collect more than the 2 % sample provided by the Streaming API. A central visualisation was created to monitor and control the individual machines' collection, as each machine was located in different physical locations around the laboratory.

Tweet-o-Meter is a visualisation of 16 real-time gauges monitoring the rate of activity of individual tweets geolocated 30 km from the administrative centre of 16 cities around the world, focusing on New York, London, Paris, Munich, Tokyo, Moscow, Sydney, Toronto, San Francisco, Barcelona and Oslo (Fig. 13.7). The first candidate cities were chosen based on the first line of chorus to the song "Pop Musik" by band M ("New York, London, Paris, Munich, Everybody talk about pop musik.") This served to ground the project into popular culture to attract users to view and use the site. The visualisation also serves a secondary purpose as a visual indicator of multiple processes on a given machine's output over time. Each gauge monitors the aggregated rate of tweets, or tweets per minute (TPM), in a given city, and the value is updated in real time.

Having multiple processes, mining data on separate physical hardware yielded approximately 12 million tweets over 3 days of collection (Weekend Friday–Monday) for a 30 km radius around London. Two separate categories of geocoded data were returned from the API: public tweets that were geolocated from a device capable of determining location (e.g. smartphone with GPS sensor) and tweets that have been positioned by reverse geocoding the location from a user's profile location. It is important to note that these tweets were flagged as publically available tweets by the creator which allowed the geolocation of a tweet to be shared with third-party applications. Through experimentation and various collections at differing times, it was found that that approximately 2 % of all tweets in a 30 km radius had detailed latitude/longitude coordinates associated with a tweet.

Using this, collection method allowed the system to successfully experiment by collecting responses for surveys created within SurveyMapper. By mining tweets for the hashtag #5Acts4wildlife, the 5Acts4wildlife campaign aimed to

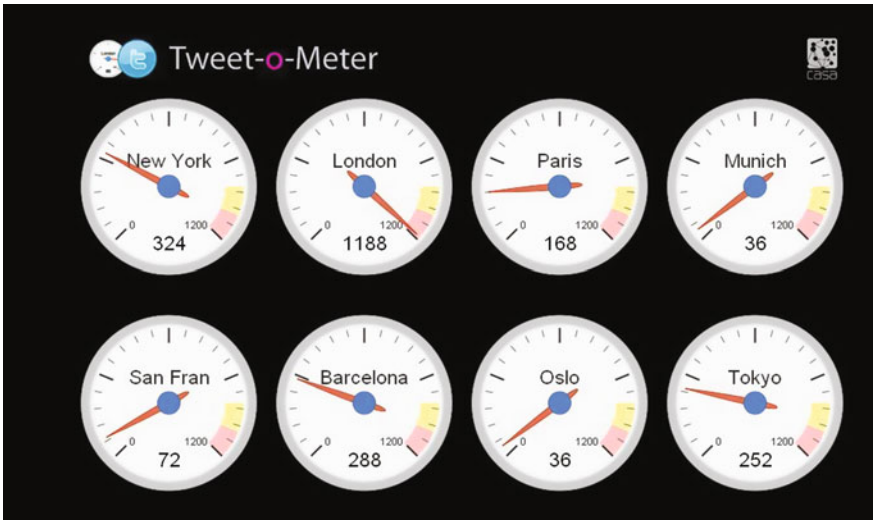

Fig. 13.7 Tweet-o-Meter real-time gauges for 8 cities. *Source* own

crowdsourced opinion on 5 campaigns that affect wildlife during a single week in January 2011. Due to the resource-intensive nature of setting up collectors to mine data for a specific survey, this process was set up manually to feed data to SurveyMapper, but recent advances in cloud computing now allow clusters of virtualised hardware to be set up automatically to mine data from Twitter and collect response data.

Tweet-o-Meter has featured on and been used by various media outlets across the world, namely CNN and Discovery Channel, Canada, during the Fukushima Earthquake in 2011 and has been used by various companies to collect social media data for visualisation and analysis. Storing the data in a central repository allowed social scientists within the laboratory to create new visualisations based on the archived city data. One such project set out to create a set of "New Cities Landscapes" which highlighted the landscape of the peaks and troughs of these hidden cities using the geolocated Twitter data. The team also collected and analysed data for the mobile phone network, EE, for analysis of the 4G mobile connectivity roll-out. The data for 10 major cities across the UK and the top 250 trending words in each city were extracted and passed on to a digital artist who created an infographic and set of visualisations for art galleries in each city.

Collecting and archiving approximately 5 million tweets a day to power the Tweet-o-Meter dials and subsequent research uses significant hardware resources. The collection of Twitter data, including relevant metadata for each tweet, for the 2 weeks of the London 2012 Olympics Games for the 22 separate venues, utilises approximately 1.5 TB of hard drive space. Therefore, to provide an archive of 4 years' worth of data for each of the 16 cities became unsustainable. At present, the system discards data 48 h after collection by the Tweet-o-Meter collectors. This ensures that space on the physical machines is not overloaded, but at the same time allows researchers to recall data within the 16 cities in the eventuality of a major incident happening within the collection boundary.

13.4 Big Data Toolkit

Researchers often rely on the skills of other researchers in different disciplines to help answer questions that are important to their own research. Researchers normally contact data suppliers and are given access to interesting datasets or discover a set of data that have been released on an Open Data store, such as the GLA London DataStore (GLA 2011). However, many of the issues of analysing the data still remain the same.

"Imagine you had a massive computer database that contained all possible measurements that could ever be made over the entire span of all space and time. You could query it with any question and it would deliver the result instantaneously. All big data is merely a subset of this the biggest data that could ever exist. What would your project ask it?" (Ramalingam 2013).

Providing a generalised toolkit for social scientists, not only to collect the data but also to analyse the data from the different services, will empower them to ask questions of the social media output without the need to learn complex APIs or build bespoke tools to gather data.

Many of the social media services provide different APIs to access the data each with differing server technologies. For example, some APIs rely on simple authentication, username and password or API keys, to access data while others require complex handshakes to be performed, oAuth2, for example. Technically proficient researchers are able to write custom applications to collect the data from these services, but lack the skill in analysing and visualising the data. Conversely, spatial, geographical and social scientists are able to analyse and visualise the data to draw conclusions about use of our cities, but sometimes lack the technical expertise to acquire the data from the source. The Big Data Toolkit seeks to help researchers by providing a toolkit with a simple interface to break down these barriers to the data and allow researchers to analyse the data in varying ways (Fig. 13.8).

Cloud Computing allows users to pool vast numbers of computing resources together to build complex and dynamic systems. The emergence of cloud computing providers has increased in modern-day computing, allowing developers to leverage the vast amount of computing power available using idle computing cycles of large companies' infrastructure. In recent years, the decrease in cost of CPU time, storage and the competitive nature of these platforms have made burst cloud computing affordable. For example, a standard virtualized machine consisting of a modest CPU (2.8 Gb Quad Core) and 50 Gb of storage, networking, bandwidth and power, situated in one of the American Amazon data centres, costs approximately $0.80 per hour, whereas the same physical machine running in a local cluster would

Fig. 13.8 Big Data Toolkit collection page showing real-time collection statistics. *Source* own

cost a few thousand pounds at the outset, which does not including ongoing running costs such as machine (electric, cooling systems, etc.).

The toolkit utilises these cloud platforms to outsource computing capacity for data collection, for a nominal cost to the end-user. By creating virtualised machines in the cloud, the user can create multiple collectors, gathering data from various locations.

To test this new method of data collection, we distributed multiple collectors on Amazon's Web Service infrastructure (AWS) during the 2011 presidential debates between Barack Obama and Mitt Romney. We decided to experiment by collecting Twitter data for an entire nation, for a single specific event, capturing the maximum amount of data possible and to test the architecture of the system. A single virtual machine image of the Big Data Toolkit was created and replicated on 87 virtual machines (each running 4 regional collectors) for a period of 4 h (Fig. 13.9). A master database served as single repository of data, and each machine over time correlated its own records and merged them into the master MySQL database. A master server acted as an intermediary between the cluster on AWS and the Big Data Toolkit desktop application, providing jobs to each machine after initial installation and configuration. Each machine connected to the master server and retrieved a latitude/longitude coordinate pair, radius and a unique collector identifier to set up the collectors, which was logged by the master server for tracking purposes. This control server relayed collection statistics, current collection totals and snapshots of data collected back to the application, providing a real-time dashboard of collection statistics. A total of 692,986 geolocated tweets (approximately 8 % of all tweets sourced) were collected in the master database (Fig. 13.10), which allowed social scientists to analyse the differences in sentiment between the East and West Coast voter opinions over a 10 min rolling average throughout the 4-h period.

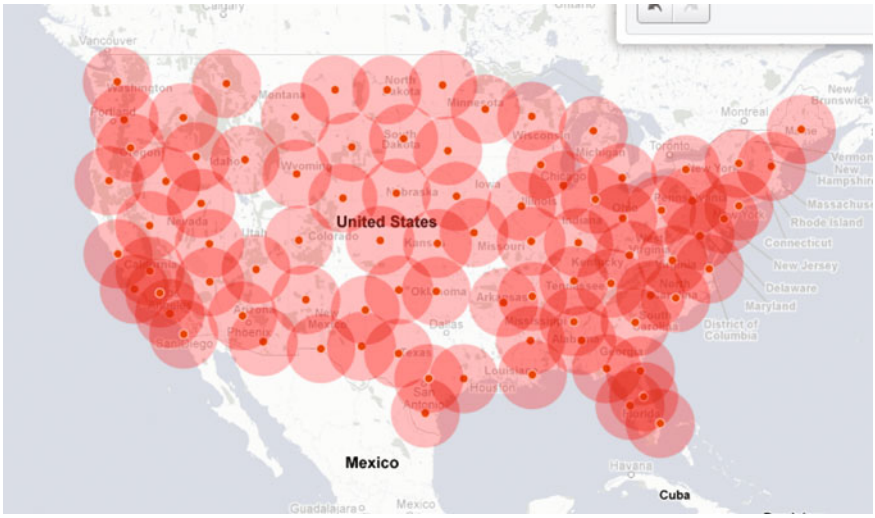

Fig. 13.9 Distributed collection experiment showing collector radius for virtual machines. *Source* own

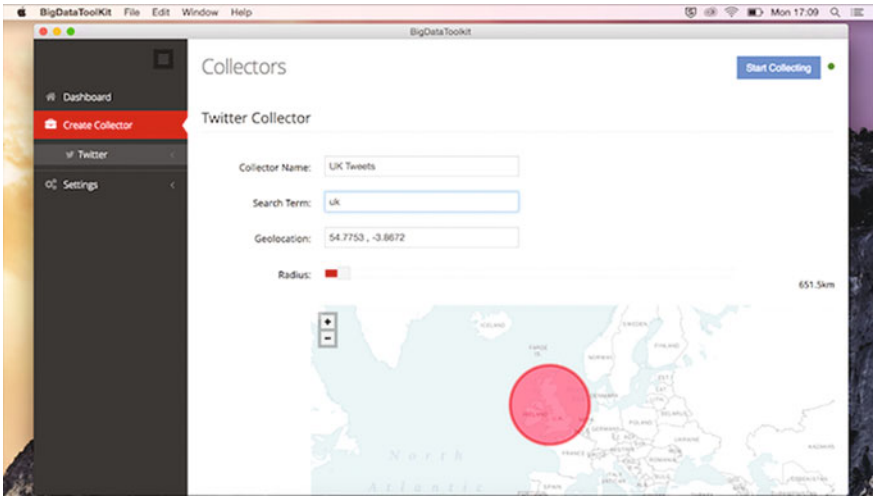

Fig. 13.10 Big Data Toolkit application—setting geographical radius for collection. *Source* own

The toolkit allows users, by combining the tools documented within the chapter, to mine and analyse various social media data in one application without having previous knowledge of the underpinning APIs. The toolkit is modular by design, and modules are added to allow users to collect data from various API (Fig. 13.8). As new services and APIs are identified, modules can then be added to the toolkit without making significant changes to base system. This data can be fed into different analytical packages, included within the toolkit, giving real-time feedback on the collection to the user. The feedback is provided via charts, data feeds, maps and dynamic interactive word clouds (Fig. 13.11) which allows the user to define new search terms, change locational regions where the data are collected or edit the data while the collection is being carried out.

The Big Data Toolkit currently provides data for a number of services that are central to UCL CASA research. The toolkit provides social media data collection for City Dashboard, a real-time dashboard of open data for various cities in the UK, and various visualisations including the London Data Table, PigeonSim, iPad Video Wall, as well as Tweet-o-Meter and SurveyMapper. The system also provides data collection services for Internet of schools, a partnership of schools that are equipped with real-time sensors and QRator, a collaboration with UCL Digital Humanities and the UCL Grant museum providing visitor engagement through digital tablets.

The toolkit stores all data within a MySQL database, which is installed locally on a user's system. All data that are collected are visualised from this central location so that the various processes and applications can share datasets. Due to the terms and conditions stipulated by some of the APIs, the sharing of data outside of the application is forbidden; therefore, by using the users' own login credentials and datastore, they can legitimately use the API while keeping within the application's terms and conditions.

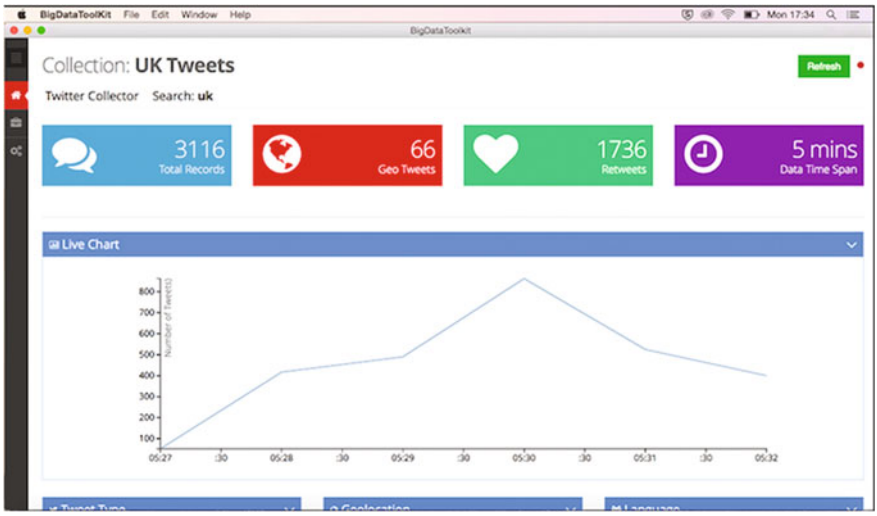

Fig. 13.11 Big Data Toolkit application—live chart of data collected within first 5 min of collection. *Source* own

Due to the modular design of the toolkit and the ability to use cloud computing to create multiple collectors to watch geographical areas, we can build a workflow to mine data continuously from different services and compare and analyse the data in real time. This provides social scientists with a complete view of all social media output for multiple locations and allows them access to the raw textual data for research purposes.

13.5 Conclusion

Crowdsourcing data for scientific projects has yielded interesting results for various applications. Projects such as Galaxy Zoo and Zooniverse have brought citizen science to the forefront of research and reduced the time taken to analyse large datasets, which would have not been possible using standard, algorithmic and processing techniques. Systems and techniques for social scientists are emerging, but there is a key need, while simplifying the collection and analysis and ensuring access to the raw data, to provide toolkits that do not require specialist tools. The crowdsourcing of geographically tagged social media has notable potential for the humanities, not to replace any current techniques but to add to the availability of data, "big data", that can be collected on demand, regardless of location and at short notice. It is moving towards an era of "as required" data collection with analysis in real time and removing the current need for knowledge about APIs or complex data collection systems. The backend to social science data collection is a complex

computer science problem, the front end, now thanks to systems such as the Big Data Toolkit, which is simply a new workflow in the research method database and the ability to tap into the crowd.

References

Amazon. (2011). Amazon mechanical turk—artificial, artificial intelligence. https://www.mturk.com/mturk/welcome. Retrieved Nov 2013.

Demarest, M. (2011). Data overload threatens with rise of smart tech and real-time sensors FCW. http://fcw.com/articles/2011/02/28/comment-marc-demarest-sensor-data.aspx. Retrieved Mar 2011.

GLA. (2011). London datastore. http://data.london.gov.uk. Retrieved Aug 2011.

Greiner, (2007). Putting Canada on the map. http://www.theglobeandmail.com/technology/putting-canada-on-the-map/article1092101, Retrieved Dec 2014.

Howe, J. (2006a). Crowdsourcing: A definition. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html. Retrieved Sept 2011.

Howe, J. (2006b). The rise of crowdsourcing. http://www.wired.com/wired/archive/14.06/crowds.html. Retrieved Dec 2013.

Hudson-Smith, A., Batty, M., Crooks, A., & Milton, R. (2009). Mapping for the masses: Accessing web 2.0 through crowdsourcing. *Social Science Computer Review, 27*, 524.

Marsh, (2014). #UKsnow Map. http://uksnowmap.com. Retrieved Jan 2015.

Met Office. (2011). The big freeze—Nov-Dec 2010. http://www.metoffice.gov.uk/about-us/who/how/case-studies/big-freeze. Retrieved July 2015.

Ramalingam, B. (2013). Lies, damned lies and big data | aid on the edge of chaos. http://aidontheedge.info/2013/02/01/lies-damned-lies-and-big-data. Retrieved Nov 2013.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation*, *67*(2), 214–237. doi: 10.1108/00220411111109449.

Sloan Digital Sky Survey. (2010). Galaxy Zoo. http://www.galaxyzoo.org. Retrieved July 2011.

Stone, B. (2006). Introducing the Twitter API. https://blog.twitter.com/2006/introducing-twitter-api. Retrieved Jan 2014.

Zooniverse. (2009). Zooniverse—real science online. https://www.zooniverse.org/. Retrieved Jan 2015.