Crowd Sourced Data for the Social Sciences: Web Based Services and Real-Time Geographic Surveys

Andrew Hudson-Smith, Richard Milton, Andrew Crooks and Mike Batty
Centre for Advanced Spatial Analysis, University College London
Email: asmith@geog.ucl.ac.uk

Abstract. Tools deployed by the social scientist are both qualitative and quantitative: analysis is based on primary data such as surveys, questionnaires and interviews, or on secondary data associated with statistics collected and collated by primary agencies (ESRC, 2009). In the last two years, there has been a distinct sea change in our ability to collect and mine both qualitative and quantitative data via web-based services. The rise of sites utilising Web 2.0 technologies such as FaceBook, YouTube, Vimeo and MySpace, to name but a few, combined with tools developed specifically as part of the remit of the National Centre for e-Social Science (NCeSS), are leading the field towards innovative and rapid methods for the collection and visualisation of data. We examine these new methods for data collection and take a look at why our system is not Grid-enabled, instead embracing the much wider environment of web-based services and rapid development cycles.

Generative social science is widely regarded as one of the grand challenges of the social sciences. As such through the ESRC-funded GENeSIS and formally Geographic Virtual Environments (GeoVUE) projects at the National Centre for e-Social Science and located in CASA, we have been releasing various software packages to make mapping accessible to the wider non-scientific community. The express aim of GeoVUE was to develop new kinds of virtual urban environments (VUEs) through which users could participate in furthering their understanding of cities and related human/social geographical systems as well as engaging in other forms of participation. The GeoVUE project was developed during a period of rapid change in the area of geographical visualisation, data collection, and data use. As such, strategic changes were made to the project, moving it away from the initial proposal to utilise GRID-based computing towards developing and focusing on the use of Web 2.0 services.

Our first notable software release in 2005, known as GMapCreator, was designed for individual users who wanted a quick, simple and perhaps more importantly free tool,
enabling thematic data to be overlaid on top of Google Maps. The stand alone software was aimed to be simple as possible with the ability to load a .shp or .csv file and click publish according to a chosen level of zoom, creating a html page and folder containing tiled images ready to view.

The introduction of Google Maps in 2005 changed the way the public and academia alike could access high-resolution base data on a global scale. Its release eased the then problematic issues regarding copyright on using street level mapping and aerial imagery over the Internet and paved the way for GMapCreator. Other web-based maps had existed prior to this, for example www.streetmap.co.uk, which covered the UK, but with Google’s map a substantial amount of the world was covered with both a street layout and a satellite view. Google’s approach to mapping is without question innovative, particularly in its use of an API (application programming interface), which allows people to use Google Maps on their own websites with a high level of flexibility. Data could finally be mixed and matched, coining the phrase ‘mashup’, where data from different sources is combined onto one map. The Javascript API which Google made public was simple enough for people with little or no programming experience to add clickable markers and lines to the maps, but was very limited in the amount of data it could handle.

Our initial experiments with thematic data overlays were based around web services that created the tiles as they were requested. However it was realised that maps could be more efficiently built by pre-creating tiles and storing them as image files on a web server. This has the advantage of reducing everything to simple file transfers between the web server and the client’s browser which is running Google Maps. With all the tiles pre-created, there is no processing element on the server, and as such it is possible to handle a high numbers of users with the additional bonus of allowing anybody with the ability to upload files to a web server can create Google Maps sites with the software. There is no complex server configuration or dependency on a particular server type as the entire system operates via file transfers. The creation of tile based imagery has an additional benefit of creating raster files direct from vector datasets. Rasterising the data while maintaining resolution allows a significant rise in performance with large datasets.
compared its vector equivalent while ensuring the data cannot be mined. Data mining of vector datasets is a serious concern in terms of copyright as it significantly limits the amount of data government and private agencies are legally allowed to share. By automatically rasterising the data, GMapCreator ensures that the process of data mining is compromised, thus opening up the possibility of much wider data visualisation. While creating the map tiles via GMapCreator was automatic, sharing multiple datasets or creating a web based service around the data required a high level of manual coding. That said London Profiler, which we illustrate in Figure 2, developed a notably successful service using the GMapCreator software. One of many users, both commercial and non-commercial GMapCreator illustrates the ability to small and rapidly developed software to make a large impact in terms of data sharing and visualisation.

![London Profiler](image)

**Figure 2.** London Profiler developed using GMapCreator.

Despite an arguably slow adoption from the social sciences with regards the utilization of Grid-based technologies there has been a notable optimism amongst researchers for the simplification of workflows and the use of Grid Upperware. This optimism is typical in the use of computers in planning, geography and indeed the wider science community in general often occurring in ‘waves’ of innovation (Barrett and Leather, 1984). Each wave is often characterised by the introduction of new software and hardware into academic use on a wave of high expectation, new funding streams and optimism. These waves of innovation and optimism relate the to Gartner’s (2008) Hype Cycles, as we illustrate in Figure 3:

![Hype Cycle](image)

**Figure 3.** The Hype Cycle of Technology (Gartner, 2008).
While Gartner's interpretation is based around mass-market adoption and interest from the wider media, rather than an academic context, it is still note worthy in terms of Grid based technology and its use by the social sciences:

1. "Technology Trigger" — The first phase of a hype cycle is the "technology trigger" or breakthrough, product launch or other event that generates significant press and interest.

2. "Peak of Inflated Expectations" — In the next phase, a frenzy of publicity typically generates over-enthusiasm and unrealistic expectations. There may be some successful applications of a technology, but there are typically more failures.

3. "Trough of Disillusionment" — Technologies enter the "trough of disillusionment" because they fail to meet expectations and quickly become unfashionable. Consequently, the press usually abandons the topic and the technology.

4. "Slope of Enlightenment" — Although the press may have stopped covering the technology, some businesses continue through the "slope of enlightenment" and experiment to understand the benefits and practical application of the technology.

5. "Plateau of Productivity" — A technology reaches the "plateau of productivity" as the benefits of it become widely demonstrated and accepted. The technology becomes increasingly stable and evolves in second and third generations. The final height of the plateau varies according to whether the technology is broadly applicable or benefits only a niche market.

The perception of Grid-based computing is arguably currently in a trough of disillusionment. In terms of the social sciences there is very little in the way of Grid-enabled datasets, especially in terms of data with a spatial constraint. In a pre-Web 2.0 environment, the Internet could be seem as providing access to mostly ‘static’ information (Laszewski and Hategan, 2005), it was a ‘read only’ environment. In this environment Grid-based computing arrived with a peak of inflated expectations. Around the same time this peak led to numerous Grid related funding calls, while Web 2.0 provided a ‘read-write’ environment providing a infrastructure for rapid development and perhaps more importantly a level playing field in terms of access to data for both the academic community and the wider public. Web 3.0 is arguably moving towards a ‘read-write-execute’ scenario which offers a significant enhancement in terms of modelling and simulation scenarios, which we will come back to later. Of note is the lack of hype surrounding Web 2.0. While the term itself has become saturated, short development cycles have meant that before hype levels can reach peak an application is already out and running. As such while there is a level of risk in adopting fast moving technologies results can be achieved on a much shorter time scale than comparable Grid-enabled projects. As such we suggest that while the Grid has been in a trough of disillusionment, ever shortening development cycles and read-write technologies via Web 2.0 have a much larger current potential for e-science community as a whole.
We develop this theme by introducing our geographical/spatial toolset known collectively as ‘MapTube’. MapTube (www.maptube.org) combines the generic idea of YouTube where users can share information with the ability of GMapCreator to create thematic maps. MapTube provides a ‘place to put maps’ as we demonstrate in Figure 4 which highlights the most viewed maps currently on the MapTube site. Every map hosted on MapTube is held on an outside server, and pulled in using the XML file which is automatically created when using GMapCreator. This allows data creators to maintain ownership of the data. MapTube allows one to view and compare different datasets as a series of layers (i.e. mashups) through either the Google Map or OpenLayers interface. MapTube in essence provides a portal for geographic data produced using the GMapCreator software.

The MapTube system operates in a broadly similar manner to services such as YouTube but on the client rather than server side. As such maps are pulled into MapTube via an XML file which is automatically created whenever GMapCreator is run. Pulling data in from outside servers rather than holding them internally allows data creators to maintain ownership of data tiles and ultimately their distribution, which is seen as a nod back to data providers who are perhaps wary of the Web 2.0 environment. Data sharing and issues of copyright have been at the forefront in the shaping of our MapTube concept.

MapTube like London Profiler allows users to link into other sites with KML files and this naturally extends to the MyMaps facility which is a kind of local customisation of Google Maps. Figure 5 shows how we can relate a series of local maps created in MyMaps to wider, more professionally created data. Serious knife crime related to teenage murders has become a significant public issue in London, and from casual reporting in the daily press, it is easy to extract the locations of these crimes and to compare these against national assaults in England which were uploaded to MapTube by an independent third party user. Data on such incidents is hard to come by from public sources as the police do not provide open access to incident locations, despite it perhaps being in the public interest. As such it is left to the ‘public at large’ to fill in the gaps. In the case of teen killings, a user with the pseudonym MapMan has created a map of teenagers murdered in London since 2007. Created using MyMaps, the list has been
compiled via various websites with street names identified in related press articles. Actual positions within the streets are not likely to be accurate, but the street names themselves are. Note the map relates to all murders, not just knife related incidents. This is shown in Figure 5(a). Using MapTube, the map can be overlaid with other data sets, such as the map uploaded detailing assault using a knife or sharp objects extracted from all 2007 hospital admissions which are classified with code ICD-10 X99. The map excludes all codes that may indicate accidental injury ICD10 – W25, W26, self inflicted injury ICD10 – X78 and undetermined intent ICD10 Y28. This is shown in Figure 5(b).

a) Murders in Greater London 2007-8   b) Distribution of Assaults in England

c) Overlays of Murders onto Assaults   d) An Overlay of Regeneration Areas

Figure 5. MapTube and Google MyMaps: Local Reporting of Murders Correlated with the National Pattern of Assaults.

Figures are standardised by age per 100,000 population while the actual counts were excluded from the map due to disclosure issues involving low numbers. By overlaying the two maps you begin to get a picture of the extent of knife crime and the number of murders in London as we show in Figure 5(c). Each link is clickable for more information. If you then use MapTube to add in Regeneration Areas within London, a clear pattern between teenage murders and deprivation emerges. MapTube enables this layer to also be viewed against other indicators such as ethnic population, barriers to housing or any of the other 47 maps relating to London in the archive. The visual correlations are of significance thus illustrating how ‘professional’ and ‘amateur’ data can add real value to the sort of insights that these comparisons enable.

Building on the use and indeed the potential of public contributed data for the social sciences we embarked on widening the reach of MapTube with the capability for more effective crowdsourcing and indeed crowdcasting of data. From independent sources, in
this case TV and radio, which is broadcast media, users were invited to log onto a web page and respond to a series of questions. Responses involve specifying their location through their postcode. As MapTube is a service which serves maps to users, it has the capability of updating a data set quite frequently by scanning the data and adding any new locations, thus producing the a new map to display. In short, TV broadcasts the invitation to respond which is backed up by the TV web site and users respond to this via the web site, but the data is then uploaded to MapTube which converts it every 30 minutes, say, to a form where it can be displayed as a map. This process was first used by the BBC radio to create a mood map of the credit crunch within the United Kingdom In conjunction with the BBC Radio 4 iPM show and NewsNight (see Figure 6a), we created a series of questions – 6 in all – which users could respond to with regard to their perceptions of their personal finances during the current economic recession, popularly known as the Credit Crunch. Based on what they considered to be the singly most significant factor hurting their personal finances, participants were asked to enter the first part of their postcode (postcode sector e.g. EC1A) so their responses could be geo-tagged choosing from one of six options: mortgage or rent, fuel, food prices, holidays, other, or the credit crunch is not affecting me, as we show in Figure 6b.

a) Radio 4 iPM Web Page on the Mood Map for the Credit Crunch

b) The User Web Questionnaire

c) Early Response Distribution

d) After 23000 User Responses

Figure 6. MapTube and Crowdsourcing: The Credit Crunch Mood Map.

Between 26th April and 29th June 2008, there were 23475 responses to the survey with 48.8% of responses saying that fuel was most significant factor hurting the person the most about the credit crunch. However there was spatial variation around the country
with more respondents within Greater London saying it was either mortgage or rent, or food.

No personal information was collected and participants were reassured that their actual locations could not be identified. This was enabled through the use of postcode sector rather than the postcode unit or building address, therefore preserving data confidentiality. Each response updated the database element of the underlying shapefile with GMapCreator running in the background to create a new map which was subsequently updated on MapTube as shown in Figure 7. Over time, as more participants entered information, the map went from blank to varying shades reflecting the responses with respect to what people were worried the most about in the postcode sector. Used in conjunction with MapTube, it allows participants and other users to take other information and lay the maps on top of one other, thus adding value to the data in ways that we cannot anticipate. The system is wide open to new users creating new map layers from their own sources, and in this senses, is driven from the bottom up by whatever users consider to be significant.

The potential of this approach for gathering spatial information is enormous for social science research. For example, it could easily be used to gather other information such as fear of household burglary, the quality of primary school education, access to local health facilities and so on. Mapping the Credit Crunch represents one of the first near real-time geographic surveys of a nation’s perceptions about a specific issue. As such the time element is also of importance as each response includes a time stamp allowing the nation’s mood to be visualised in both time and space. In excess of 23000 people took part in the survey over a three-week period creating a unique and interesting dataset which is very much of its time. The Credit Crunch Map has since led to several other surveys including BBC Look East, the nightly news programme for East Anglia in association with BBC local radio, using the system to create a map of people’s perceptions to anti-social behaviour. Using a similar data entry technique, viewers of BBC Look East were asked to answer a series of questions on their views on anti-social
behaviour at a postcode district level, with the survey attracting 8000 plus respondents, Figure 8 illustrates its use as part of a news segment on BBC Look East.

![Image](image.jpg)

**Figure 8. Crowdcasting: Mapping Anti-Social Behaviour in East Anglia Using MapTube on BBC TV Look East.**

In a ‘pre-Google’ world, none of this would have been possible for the license to use the base map and aerial imagery would have been prohibitively costly and the behind-the-scenes GIS would have been undoubtedly slow and cumbersome. We consider this in many senses to be Web 2.0 in action with the focus on data collection, outreach, communication and involving the public in the whole process. We note that there are many issues in relation to the quality of data and these are to be examined. On the other hand none of this would have been possible if we had stuck to the original plan to utilise the National Grid Service as it still seems to be early days in terms of allowing public access and indeed even the basics of Grid enabling spatial data. MapTube has also been used to gauge public option on the Congestion Charge by BBC North and is currently at the time of gathering data on the recession in association with BBC North East.

There has been an increased awareness of the power of geographical information over the last few years by the public at large. With this increased awareness has come the rise of volunteered geographic information (Goodchild, 2007), crowd sourcing (Howe, 2008) and Neogeography (Eisnor, 2006), amongst many other newly emerging terms linked to the geographic profession. Prior to the launch of Google Maps and other online mapping services, the ability for displaying and collecting fine scale and extensive data would have not been possible.

While it would be easy to hail these new techniques as a revolution in data collection for the social sciences, the way ahead must be judged with caution. Indeed questions need to be asked on issues of survey manipulation, data quality and whether or not these new methodologies are undermining the core practice of social science. Social networks, combined with web services are providing perhaps the richest and most timely datasets that we as social scientists have yet to work with. The challenge is to ensure the infrastructure is in place to fully utilise and analyse them, and for this we turn to ‘The Cloud’, which is of course also known as ‘The Grid’, turning full circle on our
development cycle back to the origins of this project in highly distributed, highly
decentralised but nevertheless powerful visual computing.

**Acknowledgements**

The authors are grateful to the Economic and Social Research Council for funding through the National Centre for e-Social Science (NCeSS).

**References**


**Biography**

Andrew Hudson-Smith is Senior Research Fellow at CASA. He led the NCeSS GeoVUE project.

Richard Milton is a research fellow at CASA specialising in computer programming, who along with Hudson-Smith initiated GMap Creator and MapTube.

Andrew Crooks is a GLA Economics Research Fellow in Urban Systems at the CASA.

Mike Batty is Bartlett Professor of Planning at UCL and the director of CASA.