How Virtual Are Virtual Methods?

A. Phippen
University of Plymouth
andy@jack.see.plymouth.ac.uk

The impact of the World Wide Web (which for the remainder of this piece to be referred to simply as "the web") is undoubtedly as significant as it was unpredictable. In some case we can see entirely new forms of social interaction – from small, initial virtual communities like those Rheingold referred to in his seminal work (Rheingold 1997), to today's massive, entirely virtual world such as Second Life[i]. The advent of Web 2.0[ii] sites such *MySpace*, *Bebo*, and *Flickr* and their relative population sizes (a recent measure claimed 25 million users of Bebo and over 70 million users of MySpace[iii]) show how people wish to use the web to facilitate social interactions, to form communities and to share experiences with people whom they have never physically met. Blogging provides the facilities for anyone with a web connection to expose their inner most thoughts to the world (and in a lot of cases, they seem very happy to do so!). While the immediate reaction of many to blogs is that they serve as a channel for narcissists, when one begins to examine the language and behaviour in blogs, one realises while the reasons for blogging may be broad and varied (from marketing and transparency within businesses to reflections on everyday life in a changing word for individuals) there is commonality in that they exist as a means of communication for people who wish to reach beyond their geographical constraints. These online diaries of everyday life are read and analysed by researchers who can gain fascinating insights into current thinking regarding all manner of issues.

Therefore, it is unsurprising to see a large body of work within the social sciences examining how people interact online, how people use the web, and how researchers might gain access to the wealth of data available in online environments. However, in this article I propose that while there is a significant body of social research that considers the online world, the methods used to research such things are what one might describe as traditional in nature. If we consider key texts, from early examples such as Jones' *Doing Internet Research* (Jones 1999) to Hine's excellent *Virtual Methods* (Hine 2005) there is a wealth of knowledge in understanding how the Internet can be used for research, and how we might understand the nature of social interaction online. However, in considering the methods discussed, there is little methodologically speaking that goes beyond the conventional within the social sciences. We see many occurrences of online focus groups, online interviews and online surveys, but very little of what one might consider novel from the methodological viewpoint.

From my own experiences I have, on a number of occasions, presented to a sociological audience regarding the use of technology in helping to understanding online behaviour. Every occasion followed the same pattern – introductory material and illustrative examples were met with interest. But once I started to examine how the technology was exploited to produce the data, the audience disengaged. On one occasion a research student of mine, who was in the audience, observed that as a slide with a protocol description appeared, note books were put down and arms were folded. I would suggest that a barrier existed between the audience and the material – while there was no change in the complexity of the material, the modification in language presents the notion that this is not longer "their" field – its technical and jargon heavy, and its not for them. I should stress that the response is not exclusive to the sociological field. However, this is the field which has the most to lose from not engaging with and exploiting the technology that underpins the online environments with which they research.

One might ask, what is wrong with applying traditional methods to online context? Why shouldn't established methods be used to consider this new social context? I absolutely agree. The work carried out in examining online life and the use of the web in methodological innovation is varied, extremely interesting and produces relevant and useful research. However, I also feel that by not embracing technology social researchers are missing a great opportunity to add another layer to the tools they use to measure behaviour and interaction. The day to day running of a web server generates a wealth of data about how people have

used the sites which run upon them, where they came from (virtually speaking), and what they did while they were there. Spidering and robot software roams the web indexing *any* document they find to build huge archives of searchable data against any query the researcher wishes to propose. As well as being a new environment to research, the web presents the researcher with new opportunities – by exploiting underlying technology and the data it produces, the research has the potential to non-intrusively explore how people interact online and to precisely measure levels of interaction.

For example, let us take the web based survey - perhaps the only technology driven technique that has mass acceptance among social researchers. It is an efficient means of rapid data collection and offers the opportunity for fast dissemination. The WebSM website[iv], a collective assembling and indexing research about the web based survey, boast more than 2,000 bibliographic references on the subject since 1998. However, closer examination of the types of paper presented show that the majority follow common themes held in traditional survey methods (sampling, measurement, questionnaire design, non-response, etc.). Even the section dedicated to "Technology" focuses on what one might refer to as front end technology – the software used to construct a survey, how to present the survey, interface techniques to present ones questions. Papers that go deeper into the technology (e.g. how one might use the underlying technologies to track respondent behaviour, to measure dropout rate, to see how many people click through from an invitation to participate to the actual survey, etc.) are few and far between. While I certainly would not want to be disparaging about the usefulness of articles other than those that examine the potential for technology to move the web based survey method forward, I do feel the extremely small proportion of such articles does go some way to justifying concerns about the enthusiasm of social researchers to embrace technology for anything other than facilitation.

I would acknowledge, however, that I do not think this is simply a case of sociologists or social researchers closing their minds to new methodological opportunities. I have worked in the past as both an IT developer and IT researcher, and the language of the discipline does seem to be deliberately exclusive. IT people will converse in three letter abbreviations (WWW, TCP, FTP??) and have a pride in the exclusion such dialogue creates. An infamous illustration of this is the Real Programmers mantra[v], which is still often quoted with a sense of pride by someone who has just implemented a particularly difficult to understand piece of program code. Such people are also rarely concerned with the social relevance of their work – I have lost count of the number of times fellow IT academics and professionals have asked me "why do you bother with that woolly human factors stuff?". And at the start of every new academic year I have to spend a period of time explaining to IT students why they have to do a module in the final stage of their degree programmes regarding the Information Society.

But if we were to try to join these two disparate disciplines together and consider their application in a methodologically innovative there is potential to do some truly great work in understanding the nature of the online world. While I do not intend to turn this piece into a tutorial on methodological innovation using web technologies, a couple of examples go some way to demonstrating the potential.

A very interesting, if not hugely rigorous, piece of work posted on a blog mid 2006 demonstrated how using a search engine to perform content analysis on a major data set (referred to by some as *Google Hacking* a site) can elicit some interesting results. The blogger performed some fairly simple queries constrained to the MySpace website, in order to examine the type of person who posts to the site[vi]. Highlights of this brief "experiment" include 9620 occurrences of "I want to kill myself" and 134,000 hits for "I hate my life" (although a quick query to Google while I write this piece shows that the number of "I want to kill myself" quotes has risen to 51,400 and "I hate my life" now numbers 1,050,000). While the work does not demonstrate anything that might not be expected from conventional wisdom, it does show two things very clearly. Firstly, social networking sites such as MySpace provide huge volumes of data about today's youth, and secondly, with unstructured data of that scale traditional content analysis methods are very difficult to use - however, a search engine, when used beyond its basic functions, can be extremely effective.

My second example differs in that it demonstrates the usefulness of the data generated by web servers with every request they deal with. A little known (outside of the technical field) aspect of web browsing is that every time a link is click, or a web resource request is typed into an address bar, the server responsible for dealing with that request writes the details to a *log file*. The entry in the file will record data such as the originating IP address[vii], the resource requested, the time and date of the request, and the referrer of the request (i.e. the site or web page the click came from). While as a single data artefact, this information is not hugely useful, when it is compiled into a large collection, it starts to become very interesting. For example, a number of entries from the same IP address around a close time period show the movement of a single user around a website – what pages they viewed, how long they spent on each page and when they left.

The analysis of web server log files (generally referred to as *web analytics*) has been prevalent in the commercial world for a number of years. Obviously, if a company is investing significant sums of money into a web based marketing campaign, they want to know whether it has been successful. Analytics provides the means to understand how people entered the website (i.e. did they come from an advertising site or paid affiliate), whether they spent a long period of time on the site, and most importantly, whether the spent any money. When log file analysis is compounded with identifiers that are passed from a specific browser on every visit (the technical term for these is *cookies* and they are passed to and from sites all of the time, without you knowing) and mixed with profile data stored in database on the site, there is a wealth of information available to the service provider. This is why Amazon knows who you are and what you'd like to buy when you login!

Now let us consider a scenario in a more social scenario. The Mediterranean Voices project[viii] was a European project to produce "an ethnography of the Southern Mediterranean" (Phippen 2004). 12 partners across the region collected multimedia resources that had associated meta-data coded based upon attributes such as location, language of description, type of resource (i.e. video, audio, image) and themes (such as the person, work, worship and play). The resources were then uploaded to a web site that provided an interface to these resources. One of the project aims was to demonstrate common themes that exist across disparate locations in the region and get people to explore their commonality through these themes. These themes highlighted the fact that, while the various locations and cultures of the southern Mediterranean may differ in appearance, they share common aspects of life (for example, Work, Worship and Play). A realisation of this commonality may contribute toward addressing cultural conflicts (the classic case within Mediterranean Voices being the divide between Nicosia North and Nicosia South – at the time it was very difficult to physically travel between the two). Through the application of analytics to the server's log file data, website database and IP address resolution techniques[ix], this aim could be clearly demonstrated to be successful. Visitors would start exploring resources at a specific location, and would "move" to different locations based upon these common themes.

As a final illustration, but not a fully developed example, consider the use of analytics to the web based survey researcher. A consistent problem with survey based research is response rate – who has received the invitation to participate, who has looked at the survey and decided not to participate, how many people who have seen the survey actually return it? In a web based environment, with some technical knowledge, all of these questions can be answered. Using an embedded image or similar in the email invitation, the log files can show how many people have read the email. If they click through to the survey from the email, an entry will appear requesting the survey page. By matching IP address (and if one wishes to be more precise, the cookie) from the email being opened to the survey page request, we can show an individual has read the email and gone to the survey page. Finally, the submission of the form will also result in an entry in the log file. Therefore, we now have a record of a respondent reading the invitation, viewing the survey, and submitting it. In instances where a respondent has viewed the survey but not submitted, this too will be reflected in the log file (showing a page view but not a form submission).

Hopefully the above discussion illustrates the potential for the exploitation of technology in social research. Do I consider the approaches to be better than traditional methods in

examining online life? Certainly not. However, I would suggest that these methods could compliment the more traditional approaches and provide evidence that could support data collected from other sources. For example, interviews with subjects regarding their internet use could be considered against their analytic profile to demonstrate whether their perceived behaviour is reflected in their actual behaviour. It should be stated that what is presented here is very much the tip of the iceberg regarding the potential for web technologies in social research. Analytics, in particular, has numerous other techniques (packet sniffing, third party cookies, web bugs, tag based browser sniffing, IP address resolution, etc.) which have potential to become useful tools to the online social researcher. They do raise some interesting ethical discussion about how far technology can be exploited before we start to infringe on the rights of the subject. While ethical considerations could fill a whole other article, I raise a couple of points for consideration. In these days of spam and phishing people are becoming very protective of their online identities (for example, their email addresses) and can become incensed at unsolicited email. And how many people are aware that every mouse click on a browser is recorded on a server, and that Google, for example, has an association between their browser/profile and every search they have ever made? In such cases, is it ethical the use this data to evaluate their behaviour/beliefs, etc.?

In concluding this piece, I would urge social researchers examining the online world to consider the potential of technologically centric methods in supporting their work. Talk with web technologists – you may realise they are equally scared of social research methods, but you will share an enthusiasm for online life, and the massive impact these technologies are making upon our society. I acknowledge that it may seem like the language and techniques are beyond the reach of non-technical researchers, but I would stress that web technologies are not difficult to understand once one has penetrated the language barrier. Virtually every academic browses the web, and therefore has foundation knowledge upon which to build their understanding of how the technologies work. And the technical detail should hold no fears for the sorts of people that can work the intricacies of SPSS or appreciate the curiosities of analysis packages such as N6!

i http://secondlife.com/. A virtual world built and owned by it “inhabitants”, of whom there are almost 4 million
ii Web 2.0 is a term given to the emerging collection of content driven sites such as those mentioned above that make people the contributors and facilitators, rather than consumers, of the content.
iii http://www.techcrunch.com/2006/08/08/bebo-passes-myspace-in-the-uk/
iv http://www.websm.org/
v There are many reproductions of the Real Programmers text, but it is generally originally attributed to Ed Post in Datamation, volume 29 number 7, and is reproduced around the web, for example: http://www-users.cs.york.ac.uk/susan/joke/quiche.htm
vi http://moneydick.com/wordpress/2006/04/23/science-of-myspace/
vii An IP address is the collection of four 3 digit numbers that comprise a device’s location online. Put simply it is the number used by computers to find other computers.
viii http://www.med-voices.org
ix IP addresses are generally assigned to specific regions and organisations so, while not precise, IP addresses can be used to approximate the location of the machine with which it is associated.

References

Hine, C. (editor) (2005). Virtual Methods: Issues in Social Research on the Internet. Oxford: Berg Publishers.
Jones, S. (edutir) (1999).'Doing Internet Research: Critical Issues and Methods for Examining the Net. London: Sage Publishers.
Phippen, A. (2004). 'An Evaluative Methodology for Virtual Communities using Web Analytics', Campus Wide Information Systems, vol. 21, no. 5.
Rheingold, H. (1993). The Virtual Community: Homesteading on the Electronic Frontier. Cambridge, Mass.: MIT Press