Researching Personal Information on the Public Web: Methods and Ethics[1]

David Wilkinson, Statistical Cybermetrics Research Group, University of Wolverhampton, UK. d.wilkinson@wlv.ac.uk
Mike Thelwall, Statistical Cybermetrics Research Group, University of Wolverhampton, UK. m.thelwall@wlv.ac.uk

There are many personal and social issues that are rarely discussed in public and hence are difficult to study. Recently, however, the huge uptake of blogs, forums and social network sites has created spaces in which previously private topics are publically discussed, giving a new opportunity for researchers investigating such topics. This article describes a range of simple techniques to access personal information relevant to social research questions and illustrates them with small case studies. It also discusses ethical considerations, concluding that the default position is almost the reverse of that for traditional social science research: the text authors should not be asked for consent nor informed of the participation of their texts. Normally, however, steps should be taken to ensure that text authors are anonymous in academic publications even when their texts and identities are already public.

Keywords: personal information, web research methods, web research ethics

Introduction

Researching personal topics, such as depression, relationship breakdowns, substance abuse and friendship, has traditionally been time-consuming and therefore expensive. Because there has been no readily-available data source comparable to a national census for most personal issues, until recently there seems to have been no methodological alternative to interviews and questionnaires. These both also have ethical issues, as human subjects research, and hence require additional researcher time and care to implement (e.g., Heath, Brooks, Cleaver, & Ireland, 2009). It seems, in theory, that the web could change this because personal issues may be found online in various places, from specialist forums (e.g., loveforum.net, cancerforums.net) to personal blogs. This may create the possibility to investigate even sensitive topics online, should it be possible to develop methods to identify appropriate texts and to deal with the theoretical implications of using a web convenience sample in research.

Many people now put personal data about themselves and their lives in the public domain through social network site (SNS) or blog profiles. A typical SNS profile may contain name, age, gender and approximate geographic location as well as selected hobbies, interests, attitudes and opinions. Whilst many SNSs have privacy settings to prevent non-Friends from accessing this information (except perhaps for people in a local network), significant numbers in major sites like MySpace, YouTube and Twitter have completely public profiles, whether from members' personal choice or by members accepting the default privacy settings at the time of joining. In theory, this gives academics and market researchers unparalleled access to mass personal data with which topics like homophily (Thelwall, 2009; Yuan & Gay, 2006), depression (Goh et al., 2009), brand perception (Jansen, Zhang, Sobel, & Chowdury, 2009) and personal taste (Liu, 2007; Liu, Maes, & Davenport, 2006) can be investigated relatively cheaply and easily.

The public web, i.e., web pages that are accessible to typical web users without any password protection, also contains a mass of informal text-based public *communication* in forums, chatrooms, blogs and SNS comments. This varies from intimate love messages between partners to public debates. Access to these texts gives new opportunities to investigate communication itself (Danet & Herring, 2007) as well as communication-related

[1] Wilkinson, D. & Thelwall, M. (2011). Researching personal information on the public Web: Methods and ethics, *Social Science Computer Review*, 29(4), 387-401.

issues like politics (Adamic & Glance, 2005) or commonly discussed topics like popular books (Gruhl, Guha, Kumar, Novak, & Tomkins, 2005). When communication occurs within friendship networks or other informal environments, such as with MySpace, YouTube and Bebo comments, and with Twitter to some extent, this seems to create a situation where personal information is particularly likely to be revealed and to be available to researchers.

This article (a) demonstrates that some research into personal information and communication is possible using only a web browser, (b) discusses some methods that involve specialist software, and (c) examines ethical considerations for the use of the resultant data. Whilst most previous publications discussing online research methods and ethics focus on a specific research topic or general methods for researching an area of the web, this is apparently the first contribution to discuss methods and ethics from the perspective of personal information. Hence it is able to offer a more general perspective on this important theme. The most similar previous paper (Hookway, 2008) has some overlap but takes a more general perspective on the type of information analysed combined with a narrower perspective on methods, focusing on qualitative blog research.

The web as a data source

The potential of the web for research has been widely acknowledged (Almind & Ingwersen, 1997; Hine, 2000; Rogers, 2005; Thelwall & Wouters, 2005). Most obviously, the web itself is extensively researched by social scientists and humanities researchers to see how and why it is used, typically with a particular topic of interest such as online politics (Pini, Brown, & Previte, 2004), globalisation in the online spread of jokes (Shifman & Thelwall, 2009), health web sites (Cui, 1999), and gender in web communication (Cressor, Gunn, & Balme, 2001; Herring & Paolillo, 2006). Moreover, there are online research methods that may be used to investigate web-related issues or may be used for convenience to investigate other topics. These methods include virtual ethnography (Hine, 2000) and web-based surveys (Couper, 2000). Finally, there are methods that passively use the web as a data source to investigate offline issues or issues for which the online component is either irrelevant or a relatively minor factor. Examples of the latter include research into trends in happiness in society by analysing the sentiment expressed in a large collection of web texts (Dodds & Danforth, in press; Kramer, 2010), research into public fears about science via large-scale blog filtering for relevant posts (Thelwall & Prabowo, 2007), research into personal morality by analysing blogs for reflectivity (Hookway, 2008), and research into social networks via data in social network sites (Ackland, 2009). The combination of large quantities of freely-available text and computerised methods of analysis (e.g., Hopkins & King, 2010) promises to be particularly potent in future social sciences research.

This article is primarily concerned with the last type of research discussed above: using the web as a data source for non-web issues. In addition to the problems of collecting appropriate data and ethical considerations, both of which are discussed below, there are also general sampling issues. If the web is used to gain data about primarily offline factors then whichever method is used is likely to have significant sampling biases. In particular, people who do not use the web are almost certain to be excluded from the sampling frame. Moreover, even web users are likely to also be excluded if they do not publish anything on the web, for example through a web site, a personal blog or social network site profile, or by commenting on others' web sites. Hence, for instance, a research method gathering data from blogs would have as its basic sampling frame web users who wrote or commented on blogs. If the project aim was to track happiness in society (Dodds & Danforth, in press) then it would need to justify the extrapolation from bloggers to wider society. For some projects this might be justified but for others extrapolation might be dangerous: for instance studies into technophobia should not use a sample of bloggers. The sample bias discussed here is in addition to any bias caused by the method itself.

Although the above discussion shows that web research can have serious limitations, it seems that few social research methods have no sampling biases (Hookway, 2008) and many enjoy widespread use despite major sampling limitations. Most qualitative research

(Creswell, 2003), for instance, uses samples that are too small to generate statistical conclusions but make valid research contributions, typically by generating insights rather than validating hypotheses. In addition, snowball sampling (i.e., chain referral sampling) is particularly subject to bias because subjects are chosen through knowing other subjects (Biernacki & Waldorf, 1981), which is the opposite of best practice for generating random samples. Nevertheless it is particularly useful for investigating marginalised and criminalised populations (e.g., Browne, 2005). In summary, web sampling biases limit the power of web research, but do not invalidate it, and mean that any conclusions drawn should always be cautious.

Methods overview

Three broad methods are described here for obtaining personal information from the web. The first two are relatively simple: composing appropriate searches for general or specialised search engines and manually filtering the results to get relevant matches. No specialist software is needed for this but filtering the results can be time-consuming if they contain many incorrect matches or much spam. The third method is to use specialist software, some of which is free on the web. This approach is essentially equivalent to the first but is more powerful and can save some human labour.

Method 1: Tailored queries in general search engines

A simple way to get insights into a topic is by composing appropriate queries and submitting them to a commercial search engine. This produces a sample of pages from the public web that could be investigated for insights into the topic. Search engines only cover the publicly indexable web pages that are not password protected and (mainly) which can be found by following chains of hyperlinks (Lawrence & Giles, 1999). This method is perhaps useful for gaining quick insights but may suffer from the heterogeneity of web publishing and it seems better to search specific web sites or types of web site that are rich in user-generated content. This can be achieved by modifying the queries to match individual SNSs, blogs or forums.

SNSs tend to have search facilities that are unsuitable for researchers because their goal is to find likely Friends. For instance, MySpace allows searching by name or e-mail address but does not have detailed searches for profile information. The public part of any web site can be fully searched, however, using a commercial engine like Google, Bing or Yahoo! by appending the keyword command `site:` followed by the appropriate web site domain name. This keyword ensures that the results are all from the specified web site and is a useful device to investigate individual sites. This type of query could be used to search for topics, attributes or even communication forms. Whilst this method is simple and widely applicable, several factors should be recalled when employing it.

- Commercial search engines do not guarantee exhaustive coverage of any web site and so the search results are drawn from an unknown proportion of the public part of the web site.
- The problem above is exacerbated by search engines tending to hide results that are “locally similar” to another result in the set (Benedict & Thomas, 2003; Thelwall, 2008; Uyar, 2009). Here, “locally similar” means that terms near to the searched keywords are mostly the same in both documents.
- In practice, the construction of appropriate keyword searches is difficult. The perfect search would return all web pages relevant to the research topic and no irrelevant pages (i.e., 100% precision and 100% recall in information retrieval terminology). This is unlikely to be achievable due to traditional information-seeking issues like term ambiguity, polysemy, synonymy, and spam (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001). The construction of an optimal keyword search therefore requires traditional librarianship skills, including an understanding of Boolean and advanced search facilities of commercial search engines. For instance, it is useful to know that `–term` added to a search excludes all pages matching the term

and putting a word in quotes may be necessary to force an exact match. In some cases it may be impossible to construct a satisfactory search.

Once optimised, the searches are likely to produce pages of results that need a final human filtering stage to remove spurious matches. The end result is a set of relevant profile pages, blog posts or other web pages. The exact analysis performed on this set depends upon the research question but a useful generic technique is content analysis: picking a set of relevant classes, describing them and then counting how many of a random sample of pages from the set fit each category (Neuendorf, 2002).

The site search method can be applied to any SNS, blog or other web site that is public and has its own domain name or uniquely identifying domain name and path. For example a MySpace profile URL is http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendid=151968769 and so profiles can be searched using a Google site-specific search for the MySpace profile URL domain name: profile.myspace.com. For example, the query `site:profile.myspace.com Northampton` matches MySpace public profiles containing the word Northampton. Note that MySpace profiles can also have other naming structures, however, such as www.myspace.com/mikethelwall. At the time of writing the following web sites, as well as many others, could be searched using the method above. Individual forums can also be searched with a similar technique, but these tend to be local or special purpose (e.g., forum.theargus.co.uk) rather than combined together into a single site, and hence are less easy to search on a large scale.

- SNSs: MySpace, Bebo, hi5. Facebook can also be searched but the results seem to exclude all personal member profiles
- Application SNSs: web sites with SNS functions but serving other purposes: YouTube, Flickr
- Blogs: Blogger, Blogspot, WordPress
- Microblogs: Twitter
- Forums: forums.myspace.com, ukbusinessforums.co.uk

Example: Relationship breakdowns

This section gives a brief case study to illustrate the above methods. The breakdown of a relationship is normally a traumatic but important event in a person’s life. As an event is associated with strong emotions that change over time it would be difficult to research retrospectively and it would be an awkward subject for an interview with an unknown researcher. It seems likely that some people would express their feelings only to their friends or to the general public in blog or SNS entries. Hence it seems possible, in theory, to identify a sample of relationship breakdown discussions on the web. Whilst topic-specific environments like alt.relationship.breakdown may be useful an alternative source, blogs and SNSs may be less affected by group emotions (Turner & Stets, 2005) and so may give more reliable results for individuals.

In order to construct searches for discussions of relationship breakdowns, a list of expressions was produced. The assumption here is that although relationship breakdowns will be described in many different ways, a list of standard expressions should be effective for identifying a reasonable number. For instance, the following seem likely: “dumped him”, “dumped me”, “dumped her”, “dumped me”, “split up”, “relationship broke down”, “left him”, “left her”, “left me”.

Each search could be tested to see how accurate it is at identifying discussions of relationship breakdowns in the SNSs and blogs listed above. A search giving too high a proportion of false matches would need to be modified either by adding extra relevant terms or extending the phrase or by subtracting irrelevant words that commonly appear. For instance, “left him” could be extended to “I left him”, therefore excluding probably irrelevant phrases like “we left him to go shopping”. Alternatively `"left me for"` could be modified to `"left me for" -jesus –dead -lyrics` because the phrase is in the popular songs “She left me for jesus” and “Left me for dead”, as well as others, and the revised query would remove many references to these songs. Some of the matches would still be incorrect, but this approach could be used to get results that do not have an unmanageable amount of spurious matches. Sometimes a phrase may also need to be forked into multiple relevant versions, such as replacing “left him” with “I left him”, “she left him” and “he left him”.

To illustrate the process on a small sample, just three separate queries were used: one for the general web, one for MySpace (by adding site:profile.myspace.com to the query) and one for a blog web site (by adding site:blogspot.com to the query).

```
"left me for" -jesus -dead –lyrics
"left me for" -jesus -dead -lyrics site:profile.myspace.com
"left me for" -jesus -dead -lyrics site:blogspot.com
```

An illustrative content analysis was conducted to show the kinds of information that can be extracted by this process. Each list of results was visited by a human coder to identify whether the pages did in fact discuss one person leaving another, the type of site, who left the author and the purpose of the post. The content analysis stopped after 100 matching pages for the web, with the other two lists not reaching this total.

Table 1. Results from the three relationship breakdown Bing searches.

Search scope	Matches	Tested	No error	About a person leaving the author
Web	920*	290	260	100
MySpace	217	217	70	32
Blogspot	110	110	73	60

*More hits were estimated to exist but only 920 were returned by Bing.

From Table 1 it is clear that there were many results overall but probably too few from MySpace and Blogspot to conduct a meaningful quantitative analysis. The numbers could be increased by combining the results with different searches, such as those mentioned above (e.g., “dumped me”, “split up”), and the numbers for Blogspot could be supplemented by other blog hosting sites like blogger.com. Nevertheless, it seems that a lot of work would be needed to get, say, a total of at least 100 valid results for MySpace.

Table 2 suggests approximate gender equality in the person discussing who left them, except perhaps for married couples, where the wife may be more likely to discuss being left by her husband.

Table 2. The person leaving the author (bracketed numbers are the share from same-sex relationships).

Search scope	Husband	Wife	Girlfriend	Boyfriend	Other
Web	24	10	34(4)	31	partner
MySpace	4	2	12(1)	13(1)	girlfriends
Blogspot	15	9	18(2)	15(2)	husbands

For the general web search the main sources were: forums (19), advice or agony sites (12), chatrooms (6), news sites (5), and poetry sites (5). The results included a wide variety of sources, however, including a personal story about an abortion on a health centre web site, an alcoholics help site, and customer comments for a motorbike shop (a woman claiming that her husband returned after she confiscated a motorbike accessory present bought from the shop.

Table 3 shows the range of different contexts in which the breakup was mentioned, and the categories are explained in more detail below. These differ quite substantially between sources, showing that the web should not be treated as a single homogeneous source of information. In the Blogspot category there were also 6 cases where the fact was mentioned as a small part of a longer story. Other reasons for writing were: to warn others against threesomes, to laugh at the ex-partner, to give an excuse for a crime, to directly request sympathy.

- Seek advice – about what to do about the partner leaving
- Brief statement – that the partner had left
- Complain – about how bad the partner was
- Tell the story – about how the break-up happened.
- To explain something – for example, why the author was sad, why they had to move house, or their career choice.

Table 3. Reasons for publicly posting relationship breakdown information.

Search scope	Seek advice	Brief statement	Complain	Tell the story	To explain something	Other
Web	26	29	13	17	2	13
MySpace	0	12	7	1	10	2
Blogspot	5	11	13	4	14	13

In summary, the example shows that it is possible to get enough data for quantitative analysis about private subjects from the web, but that extensive manual filtering may be needed to eliminate spurious matches like song lyrics, and that the typical content found is likely to differ between areas of the web. For a full-scale analysis of relationship breakdowns more work would be needed to get sufficient data, for instance by combining the results with those from other searches, as discussed above. It is likely that multiple searches would be needed in many other cases too, and the use of multiple searches would also help to minimise the chances of the results being biased by the search term selection.

Method 2: Generic blog and forum searching

A number of specialist blog and forum search engines give direct access to large numbers of blogs or forums. Their use can be more efficient than searching individual blogs using the method above. These search engines sometimes allow date-specific searches, which can also be helpful.

Blog and forum search engines have the same disadvantages as the site search. Their coverage seems less reliable and more unknown than that of commercial search engines, however, because they do not seem to report which forum or blog sites or types of blogs or forums they index.

The list below is of some blog and forum search engines available as at February 2010 but note that there were no SNS search engines and the forum list also covers bulletin board and chatroom search engines.

- Blogs: BlogPulse, Google Blog Search, Technorati, Twitter Search
- Forums: BoardReader, Omgili
- Newsgroups: Google Groups

An alternative technique for searching blogs or forums is to use Google’s inurl command by adding `inurl:forum` to a search to get results mainly from forums or `inurl:blog` to get results mainly from blogs.

Example: Visits to an amusement park

Suppose that an academic or market researcher wished to know the main talking points generated by a visit to a major UK amusement park. Some insights into this issue could be gained by identifying discussions of the parks in blogs and forums and filtering out those not from recent attendees. This is easily achieved for parks with distinctive names (e.g., Alton Towers) by simple name searches in blog and forum search engines. For more common names (e.g., Legoland) some search refining may be needed. The following searches illustrate the number of results returned by some of the above strategies for Alton Towers searches.

- Omgili: (`"Alton Towers"`) 916 discussions
- BoardReader: (`"Alton Towers"`) 1,000 hits
- Google Blog Search: (`"Alton Towers"`) 1,586 hits

- BlogPulse: (`"Alton Towers"`) 55,375 hits
- Google inurl forum search: (`"Alton Towers" inurl:forum`) 510,000 hits
- Google inurl blog search: (`"Alton Towers" inurl:blog`) 63,500 hits

Method 3: Specialist software

A number of researchers have developed specialist software to download information from individual web sites or types of web site for analysis. This approach is particularly useful when extensive personal information is needed, such as demographics, if a program can extract such information. It is possible when the web site owner does not disallow it in their terms and conditions. In some cases the information can be extracted by a specialist web crawler whereas in others there is an extra method of accessing the data provided by the web site owner, known as an applications programming interface (API). In either case a typical social scientist may either need to find an appropriate program on the web (e.g., free software at sourceforge.net or LexiURL Searcher lexiurl.wlv.ac.uk for automated YouTube, Technorati and web searches) or collaborate with a computer scientist to create it.

An example of an analysis of personal information extracted with a program is a study of homophily in MySpace (Thelwall, 2009). It found that Friends tended to have similar ages, religions and attitudes towards children but there was no evidence of gender homophily. Another computer-assisted analysis of MySpace found that Friends tended to express their preferences differently from each other (Liu, 2007). Out of the many automated blog analyses, one found evidence that patterns of discussing books online could sometimes predict future online sales (Gruhl et al., 2005).

The limitations with using automated software to extract data tend to be less marked in the dimensions of sample selection than the other two methods. This is because the coverage of the software can usually be precisely known, as it follows parameters set by the researcher. Of course, the data still has the same limitations when attempting to extrapolate to offline phenomena, however.

Ethics

Online research ethics have attracted interest due to the emergence of new issues and the reformulation of old ones (Ess, 2009; Ess & Committee, 2002; Eynon, Schroeder, & Fry, 2009; Sharf, 1999; Tavani, 2008; Thelwall & Stuart, 2006; White, 2002) and it seems unlikely that uniform guidelines will emerge for internet research (Jankowski & van Selm, 2005). The methods discussed above have ethical implications since they deal with personal information. Many papers about internet research have focused on email questionnaires, or interactions in chat rooms or virtual reality and none have focussed on the issue of the analysis of personal data from the public web. In consequence, there are no specific ethical guidelines or detailed discussions for the large-scale analysis of personal data. The main issues are the classical social science research concerns of informed consent, privacy, and anonymity (Heath et al., 2009, p. 21-38). The further classical concerns of confidentiality and researcher-participant interaction are not relevant, however, since passive data collection research does not involve direct interaction between participants and researchers.

Note that this section is concerned with information that is in the public web, as a collection of published web pages, rather than information that is available via the web but ephemeral, such as the utterances of participants in an online game (e.g., Roberts, Smith, & Pollock, 2004) or password-protected web pages such as message boards only visible to users (Sanders, 2005). In these cases, the researcher could be seen as a covert observer, triggering more significant concerns. In the former case the ephemeral content, whilst publicly visible for a short time, is much less like a published document than most web content. Finally, from a wider perspective social science research is important and therefore it should not be slowed down with ethics hurdles, unless necessary (Dingwall, 2006).

Informed consent: Permission to research personal information

Should researchers ask permission to use others' personal information on the public web for research? Some studies based on public web data have produced conclusions that the creators of the information may not be in agreement with (Introna & Gibbons, 2009). In many social science areas a key ethical principle is informed consent: normally the written agreement of the participants in any study about what their input can be used for. Nevertheless, the appropriate ethical procedure to be followed depends in part on how the research objects are conceived. A distinction can be made between individuals and documents as research objects. Whereas individuals tend to be protected by ethical procedures, documents can often be used without creating ethical issues, even if the research has a consciously negative impact on authors (e.g., a derogatory literary criticism). There is a precedent for conducting research based upon documents containing personal information without obtaining consent: "In clinical studies non-intrusive research such as retrospective use of existing medical records may be conducted ethically without the express consent of the individual subjects if the material is anonymised at the earliest possible stage, if there is no inconvenience or hazard to the subjects, and if the institutional review board has reviewed and agreed the research protocol" (Eysenbach & Till, 2001). Note that this precedent concerns data that is not in the public domain in any sense and is of a potentially sensitive or intimate nature and therefore needs additional protection compared to more public data (Nissenbaum, 2004).

Web-based objects like social network sites, bulletin boards and blogs are all, in principle, electronic documents (Ess & Committee, 2002). Research involving such public web documents without contacting their authors is not human subjects research (Enyon, Schroeder, & Fry, 2009) and can therefore avoid even triggering the need for consideration by university ethics committees. From a humanities perspective, Internet texts can be viewed as cultural production rather than interfaces to human subjects, with the consequent removal of the human subjects from the frame of reference altogether (Bassett & O'Riordan, 2002; Hookway, 2008; White, 2002) and so informed consent is not normally necessary. In fact, it seems that seeking informed consent can be problematic because contacting content creators goes some way towards involving them in the research, hence triggering human subjects concerns. To give an extreme example, a study of suicide-related online discussions would need to consider the situation carefully before contacting content creators for permission to use their texts. Informing subjects also has a possible negative impact on privacy because one of the recognised benefits of privacy is freedom from the feeling of being watched (Gavison, 1980), and so people might curtail their activities when the possibility is raised that they might be watched by the researcher or others. Probably, however, most people would prefer to know that they could be watched, given the choice.

Although web texts can be treated as documentary research sources or cultural artefacts, they deserve special consideration because they are less obviously public than a published book and often contain personal information. This is an issue of privacy rather than consent, however.

Access to data and privacy: Normative and natural

Should researchers have the right to others' personal information on the public web or should this be regarded as an invasion of privacy? A potential response to this is to regard the web as a collection of documents, and hence the issue is one of copyright rather than privacy – with researchers able to access the documents on a fair dealing basis (Hookway, 2008). The existence of many public blogs and social network sites containing information for friends to read has raised high-profile privacy concerns, however: if employers view a person's MySpace and decide not to employ them based upon information discovered, is this an invasion of privacy? Similarly, if a cyberstalker finds out a person's address and other information about them online, are they invading privacy? The issue here is that the person who posted the information online may not be aware of the potential for others to access it and may believe that the information is private when it is not (Tavani, 2005).

The use of implicit and explicit online personal data by businesses is common, even when it is not in the public domain, and can raise concerns about privacy and individuality (van Wel & Royakkers, 2004). For instance, organisations with extensive personal information can use it to sort or categorise individuals to target marketing or resources based upon previous experience with apparently similar people, giving them more power over individuals (Gandy, 1993). Google is an important example because it maintains a large amount of information about the public, such as search terms used and web sites visited after searching. For users of other Google services, it may also know about all web sites visited, email (if via Gmail), blogging (Blogger) and social networking (Orkut) (Zimmer, 2008). Google may use this data to target advertising, and in some cases may also give it to the authorities to be employed against the user. In these cases, the capture and storage of digital information could legitimately be seen as a form of surveillance because the objective is to manipulate the user (Zimmer, 2008). In addition, many organisations routinely capture information about web site visitors and use it in the surveillance sense of targeting personalised advertising – personalisation is an important web technology (Mobasher, Cooley, & Srivastava, 2000). Commercial uses such as these are relevant because "information gathering and processing" should be governed by the norms of the appropriate context (Nissenbaum, 2004). As a result, commercial surveillance uses of internet data helps to create norms in which less intrusive research use of internet data are more acceptable.

Academic research tends not to be a type of surveillance, in the sense that it does not directly seek to influence individuals, but it may be that the individuals would not agree with the use made of their data. Moor's (2004) theory of types of privacy is helpful here (see also Tavani, 2005): a situation is *naturally private* if a reasonable person could expect themselves or their information to be hidden from others; whereas a situation is *normatively private* if a reasonable person could expect that others would protect their privacy. An example of a naturally private situation is a person in a remote place: they could reasonably expect that whatever they did would not be observed. In contrast, a person who gave bank information to a retailer could reasonably expect the retailer to protect that information from malicious use, and so this is a *normatively private* situation. The distinction is important because when someone intrudes on a naturally private situation then this could be seen as an accident whereas when someone intrudes on a normatively private situation then this could be seen as an invasion of privacy, and legal or other redress may be appropriate. Nevertheless, drawing upon theories of privacy (Tavani, 2007), in naturally private situations a person may have an *interest* in privacy even if they do not have the (legal) right to it.

The posting of personal information in blogs, social network sites and other publicly accessible places online is one of natural rather than normative privacy because there is no reasonable expectation that others ought to be protecting such information. Whilst many users of social network sites may have misplaced expectations of privacy (Acquisti & Gross, 2006) this does not mean that they should have the right to privacy. In fact, there does not seem to be a significant movement towards normative privacy in the sense of enacting privacy laws for public data; rather, there seem to be personality differences in privacy concerns (Yao, Rice, & Wallis, 2007) and people seem to have avoided using the internet as far as possible if their concerns were significant (Metzger & Docter, 2003). In addition it seems that individuals adapt their behaviour to ward off threats, when perceived (Viégas, 2005), and apply pressure to particular web sites, such as MySpace or Facebook, when specific threats occur (boyd, 2008). In this context, loss of privacy in the sense that others (i.e., researchers) are accessing personal information is not a *violation* of privacy, but more of an *accidental* occurrence and one that the individual could not expect to have others protect them from. Hence it seems reasonable, in principle, to use personal data from the public web for research purposes.

Despite accepting that principles of natural privacy apply to public web data, it could be argued that the people concerned should be informed before their data is mined, as they could not reasonably expect to have knowledge of this information. Such a claim has been made for data mining with customer data, for example (Tavani, 2008). Data mining information on the public web is less serious, however, because the information concerned is already public, whereas information submitted to online merchants can be private in the sense of its owner affording limited access to it.

The difference between types of privacy is less important than context in Nissenbaum's (2009) theory of contextual integrity, and this could therefore potentially undermine any claims relying upon the simpler idea of normative against natural privacy. In particular, Moor's theory seems insufficient to deal with situations in which outrage is generated by actions that only breach *natural* privacy. Nissenbaum argues for moral frameworks based upon *context* to decide contentious issues rather than simple dichotomies based upon definitions of privacy or types of privacy. With reference to social network sites, Nissenbaum (2009, p. 221-230) argues that users have had unpleasant surprises due to their personal information being broadcast more widely than they deemed appropriate for context (e.g., Facebook's news feed feature) or used for purposes that they did not like (e.g., used for targeted advertising). In all the examples of public reactions discussed, however, there was a direct impact on users that caused a reaction or protests. In contrast, academic research on the same data has only a very indirect impact via the uses made of the knowledge gained. The most difficult case seems to be controversial research projects, such as those covering attitudes to abortion, because some may wish to claim a breach of contextual integrity as a defence against the research. Nevertheless, whilst this could perhaps be part of a general argument for legislation against all processing of public personal information, it is not a reasonable defence against non-intrusive research because the widespread commercial use of this data makes it (currently) contextually appropriate to use it for non-surveillance purposes.

Anonymity

Should researchers ensure that the creators of any personal information on the public web are kept anonymous from all third parties? It is normal in human subject research to assure participant anonymity as far as possible and to inform participants of any possible threat to their anonymity (Heath et al., 2009). The issue of anonymity is more complex with public internet data, however. Since only public information is being used, the creators of this information may already not be anonymous due to posting their identity, or clues to it, in the data researched. In consequence, revealing clues to the identity of an originator of some data analysed, such as their profile URL or an identifiable quote, is not breaching their anonymity but merely copying their identity from one public situation (the Web) to another (an academic article). Nevertheless, as in the case of closed-circuit television footage from public places, which is subject to laws that restrict its use by broadcast media (Taylor, 2009), this may serve to draw attention to the people involved and hence could be seen as breaching privacy. In consequence, it seems necessary to avoid including identifying information in research publications and to avoid quotes or anonymise them by paraphrasing or altering words so that they are not searchable (Ess, 2007); altering information also seems accepted practice in offline research to preserve anonymity (Heath et al., 2009, p. 35).

The need for anonymity is not universal because the authors of web texts may benefit from the publicity in some cases (Bassett & O'Riordan, 2002), or may already be public figures, such as prominent bloggers or media personalities. Researchers should be sensitive to legal frameworks and cultural norms applying to the region from which the content was posted, however (Ess & Committee, 2002).

Conclusions

The key argument in this article is that it is possible and ethical to extract personal information, opinions and attitudes from the web on a large scale for research purposes. In addition, methods have been described that make the data available using only a web browser and not specialist software. This data can be ethically researched as long as safeguards are taken to ensure that text authors are anonymous, where appropriate, and it is not normally necessary or desirable to seek informed consent.

Whilst the capability to investigate personal information provides in theory easy access to information that was previously difficult to obtain, there are two limitations. First, as with any new approach, inertia may hinder its uptake (Katz & Allen, 1982) and lack of experience with relevant problems makes ethical issues harder to assess (Enyon et al., 2009) and so studies may be more cautious. Second, and most seriously, sampling limitations make it hard to draw statistically robust conclusions. While this seems to be an issue for most social sciences data, it is probably more marked with web data. As a result the approaches described here are most recommended for initial pilot studies, when triangulation is possible, when no practical alternative is available, and when the web itself is part of the scope of a study. Pilot studies seem particularly valuable since the results may help to ensure that subsequent larger-scale research is well formulated and informed.

References

Ackland, R. (2009). Social network services as data sources and platforms for e-researching social networks. *Social Science Computer Review, 27*(4), 481-492.

Acquisti, A., & Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the Facebook. *Lecture Notes in Computer Science, 4258*, 36-58.

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. *WWW2005 blog workshop*, Retrieved May 5, 2006 from: http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf.

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation, 53*(4), 404-426.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2-43.

Bassett, E. H., & O'Riordan, K. (2002). Ethics of Internet research: Contesting the human subjects research model. *Ethics and Information Technology, 4*(3), 233-247.

Benedict, G., & Thomas, S. B. (2003). Detecting query-specific duplicate documents. *USPTO Online*, Retrieved January 6, 2009 from: http://patft.uspto.gov/netacgi/nph-Parser?Sect2001=PTO2002&Sect2002=HITOFF&u=%2002Fnetahtml%2002FPTO%2002Fsearch-adv.htm&r=2001&p=2001&f=G&l=2050&d=PTXT&S2001=6615209.PN.&OS=pn/6615209&RS=PN/6615209.

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods and research, 10*(2), 141-163.

boyd, d. (2008). Facebook's privacy trainwreck: Exposure, invasion, and social convergence. *Convergence, 14*(1), 13-20.

Browne, K. (2005). Snowball sampling: using social networks to research non-heterosexual women. *International Journal of Social Research Methodology, 8*(1), 47-60.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*(4), 464-494.

Cressor, F., Gunn, L., & Balme, H. (2001). Women's experiences of on-line e-zine publication. *Media culture and society, 23*(4), 457-473.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches (2 ed.).*Thousand Oaks, CA: Sage.

Cui, L. (1999). Rating health web sites using the principles of citation analysis: A bibliometric approach. *Journal of Medical Internet Research, 1*(1).

Danet, B., & Herring, S. C. (Eds.). (2007). *The multilingual internet: Language, culture, and communication online*. Oxford: Oxford University Press.

Dingwall, R. (2006). Confronting the anti-democrats: The unethical nature of ethical regulation in social science. *Medical Sociology Online, 1*(1), 51-58.

Dodds, P. S., & Danforth, C. M. (in press). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*.

Enyon, R., Schroeder, R., & Fry, J. (2009). New techniques in online research: Challenges for research ethics. *21st Century Society, 4*(2), 187-199.

Ess, C. (2007). Internet research ethics. In A. Joinson, K. McKenna, T. Postmes & U.-D. Reips (Eds.), *Oxford Handbook of Internet Psychology* (pp. 487-501). Oxford: Oxford University Press.

Ess, C. (2009). Floridi's philosophy of information and information ethics: Current perspectives, future directions. *The Information Society, 25*(3), 159-168.

Ess, C., & Committee, A. E. W. (2002). Ethical decision-making and Internet research. Recommendations from the aoir ethics working committee. Retrieved April 17, 2008 from: http://www.aoir.org/reports/ethics.pdf.

Eynon, R., Schroeder, R., & Fry, J. (2009). New techniques in online research: Challenges for research ethics. *Twenty-First Century Society, 4*(2), 187-199.

Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. *British Medical Journal, 323*(7321), 1103-1105.

Gandy, O. (1993). *The panoptic sort: A political economy of personal information*. Boulder, CO: Westview Press.

Gavison, R. (1980). Privacy and the Limits of the Law. *Yale Law Journal, 89*(3), 421-471.

Goh, T.-T., Huang, Y.-P., Journal, 2009, Y., 39, V., 3, I., et al. (2009). Monitoring youth depression risk in Web 2.0. *VINE, 39*(3), 192-202.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). The predictive power of online chatter. In R. L. Grossman, R. Bayardo, K. Bennett & J. Vaidya (Eds.), *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 78-87). New York, NY, USA: ACM Press.

Heath, S., Brooks, R., Cleaver, E., & Ireland, E. (2009). *Researching young people's lives*.Thousand Oaks, CA: Sage.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459.

Hine, C. (2000). *Virtual Ethnography*. London: Sage.

Hookway, N. (2008). Entering the 'blogosphere': some strategies for using blogs in social research. *Qualitative Research, 8*(1), 91-113.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229-247.

Introna, L., & Gibbons, A. (2009). Networks and Resistance: Investigating online advocacy networks as a modality for resisting state surveillance. *Surveillance & Society, 6*(3), 233-258.

Jankowski, N., & van Selm, M. (2005). Epilogue: Methodological concerns and innovations in internet research. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet* (pp. 199-207). London: Berg.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science & Technology, 60*(11), 2169-2188.

Katz, R., & Allen, T. J. (1982). Investigating the Not Invented Here (NIH) Syndrome: A look at the performance, tenure and communication patterns of 50 R&D project groups. *R&D Management, 12*(1), 7-19.

Kramer, A. D. I. (2010). An unobtrusive behavioral model of "Gross National Happiness". *Proceedings of CHI 2010*, 287-290.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400*(6740), 107-109.

Liu, H. (2007). Social network profiles as taste performances. *Journal of Computer-Mediated Communication, 13*(1), Retrieved June 5, 2008 from: http://jcmc.indiana.edu/vol2013/issue2001/liu.html.

Liu, H., Maes, P., & Davenport, G. (2006). Unraveling the taste fabric of social networks. *International Journal on Semantic Web and Information Systems, 2*(1), 42-71.

Metzger, M. J., & Docter, S. (2003). Public opinion and policy initiatives for online privacy protection. *Journal of Broadcasting & Electronic Media, 47*(3), 350-374.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on Web usage mining. *Communications of the ACM. Volume 43, Number 8,Pages, 43*(8), 142-151.

Moor, J. H. (2004). Towards a theory of privacy for the information age. In R. A. Spinello & H. T. Tavani (Eds.), *Readings in CyberEthics* (2nd ed., pp. 407-417). Sudbury, MA: Jones and Bartlett.

Neuendorf, K. (2002). *The content analysis guidebook.* London: Sage.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 17*(1), 101-139.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy and the integrity of social life*.Stanford, CA: Stanford University Press.

Pini, B., Brown, K., & Previte, J. (2004). Politics and Identity in Cyberspace: A case study of Australian women in agriculture online. *Information, Communication & Society, 7*(2), 167-184.

Roberts, L. D., Smith, L., & Pollock, C. (2004). Conducting ethical research online: Respect for individuals, identities, and the ownership of words. In H. Nemati (Ed.), *Information Security and Ethics: Concepts, Methodologies, Tools and Applications* (pp. 156-173). Hershey, PA: Information Science Publishing.

Rogers, R. (2005). *Information politics on the Web*. Massachusetts: MIT Press.

Sanders, T. (2005). Researching the online sex work community. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet* (pp. 67-79). London: Berg.

Sharf, B. F. (1999). Beyond Netiquette: The ethics of doing naturalistic discourse research on the internet. In S. G. Jones (Ed.), *Doing internet research* (pp. 243-256). Thousand Oaks CA: Sage Publications.

Shifman, L., & Thelwall, M. (2009). Assessing global diffusion with Web Memetics: The spread and evolution of a popular joke. *Journal of the American Society for Information Science and Technology, 60*(12), 2567-2576.

Tavani, H. T. (2005). Search engines, personal information, and the problem of protecting privacy in public. *International Review of Information Ethics, 3*, 40-45.

Tavani, H. T. (2007). Philosophical theories of privacy: Implications for an adequate online privacy policy. *Metaphilosophy, 38*(1), 1-22.

Tavani, H. T. (2008). Floridi's ontological theory of informational privacy: Some implications and challenges. *Ethics and Information Technology, 10*(2-3), 155-166.

Taylor, N. (2009). State surveillance and the right to privacy. *Surveillance & Society, 1*(1), 66-85.

Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology, 59*(11), 1702-1710.

Thelwall, M. (2009). Homophily in MySpace. *Journal of the American Society for Information Science and Technology, 60*(2), 219-231.

Thelwall, M., & Prabowo, R. (2007). Identifying and characterising public science-related concerns from RSS feeds. *Journal of the American Society for Information Science & Technology, 58*(3), 379-390.

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology, 57*(13), 1771-1779.

Thelwall, M., & Wouters, P. (2005). What's the deal with the web/Blogs/the next big technology: A key role for information science in e-social science research? *Lecture Notes in Computer Science, 3507*, 187-199.

Turner, J. H., & Stets, J. E. (2005). *The sociology of emotions*. Cambridge: Cambridge University Press.

Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science, 35*(4), 469-480.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology, 6*(2), 129-140.

Viégas, F. B. (2005). Bloggers' expectations of privacy and accountability: An initial survey. *Journal of Computer-Mediated Communication, 10*(3), Retrieved February 5, 2010 from: http://jcmc.indiana.edu/vol2010/issue2013/viegas.html.

White, M. (2002). Representations or people? *Ethics and Information Technology, 4*(3), 249-266.

Yao, M. Z., Rice, R. E., & Wallis, K. (2007). Predicting user concerns about online privacy. *Journal of the American Society for Information Science & Technology*, *58*(5), 710-722.

Yuan, Y. C., & Gay, G. (2006). Homophily of network ties and bonding and bridging social capital in computer-mediated distributed teams. *Journal of Computer-Mediated Communication*, *11*(4), 1062-1084.

Zimmer, M. (2008). The gaze of the perfect search engine: Google as an infrastructure of dataveillance. In A. Spink & M. Zimmer (Eds.), *Web search: Multidisciplinary perspectives* (pp. 77-99). Berlin: Springer.