Doing Blog Research: The Computational Turn

Axel Bruns and Jean Burgess

ARC Centre of Excellence for Creative Industries and Innovation

Queensland University of Technology, Australia

Blogs and other online platforms for personal writing such as LiveJournal have been of interest to researchers across the social sciences and humanities for a decade now. Although growth in the uptake of blogging has stalled somewhat since the heyday of blogs in the early 2000s, blogging continues to be a major genre of Internet-based communication. Indeed, at the same time that mass participation has moved on to *Facebook*, *Twitter*, and other more recent communication phenomena, what has been left behind by the wave of mass adoption is a slightly smaller but all the more solidly established blogosphere of engaged and committed participants. Blogs are now an accepted part of institutional, group, and personal communications strategies (Bruns and Jacobs, 2006); in style and substance, they are situated between the more static information provided by conventional Websites and Webpages and the continuous newsfeeds provided through *Facebook* and *Twitter* updates. Blogs provide a vehicle for authors (and their commenters) to think through given topics in the space of a few hundred to a few thousand words – expanding, perhaps, on shorter tweets, and possibly leading to the publication of more fully formed texts elsewhere. Additionally, they are also a very flexible medium: they readily provide the functionality to include images, audio, video, and other additional materials – as well as the fundamental tool of blogging, the hyperlink itself.

Indeed, the role of the link in blogs and blog posts should not be underestimated. Whatever the genre and topic that individual bloggers engage in, for the most part blogging is used to provide timely updates and commentary – and it is typical for such material to link both to relevant posts made by other bloggers, and to previous posts by the present author, both to background material which provides readers with further information about the blogger's current topic, and to news stories and articles which the blogger found interesting or worthy of critique. Especially where bloggers are part of a larger community of authors sharing similar interests or views (and such communities are often indicated by the presence of yet another type of link – in blogrolls, often in a sidebar on the blog site, which list the blogger's

friends or favourites), then, the reciprocal writing and linking of posts often constitutes an asynchronous, distributed conversation that unfolds over the course of days, weeks, and months.

Research into blogs is interesting for a variety of reasons, therefore. For one, a qualitative analysis of one or several blogs can reveal the cognitive and communicative processes through which individual bloggers define their online identity, position themselves in relation to fellow bloggers, frame particular themes, topics and stories, and engage with one another's points of view. It may also shed light on how such processes may differ across different communities of interest, perhaps in correlation with the different societal framing and valorisation of specific areas of interest, with the socioeconomic backgrounds of individual bloggers, or with other external or internal factors. Such qualitative research now looks back on a decade-long history (for key collections, see Gurak, *et al*., 2004; Bruns and Jacobs, 2006; also see Walker Rettberg, 2008) and has recently shifted also to specifically investigate how blogging practices differ across different cultures (Russell and Echchaibi, 2009). Other studies have also investigated the practices and motivations of bloggers in specific countries from a sociological perspective, through large-scale surveys (e.g. Schmidt, 2009). Blogs have also been directly employed within both K-12 and higher education, across many disciplines, as tools for reflexive learning and discussion (Burgess, 2006).

The Computational Turn

Over the past years, another significant approach to blog research has developed as part of what David Berry (forthcoming) has described as the "computational turn": the increased availability of tools for the semi-automatic capture and analysis of large corpuses of Web (and other) content, and the development of research methodologies across the humanities and social sciences which exploit these new opportunities, often in interdisciplinary research teams. In blog research, the thrust of such developments is twofold: the earliest such research projects (see e.g. Adamic and Glance, 2005) focussed mainly on the links found on blog sites, and used them to plot the network structure which these multiple interlinkages reveal. This is done by using Web crawler software – such as the freely available IssueCrawler (www.issuecrawler.net) – which start from a seed list of Webpages provided by the researcher, and identify and follow the links present on those pages over a number of

iterations to identify the wider network of sites which these seed sites link to. By analysing these linkage patterns, it is possible for example to identify those sites which receive the most links (and which can therefore be understood as key providers of information or opinion), those which link the most, but do not receive many links themselves (that is, sites which sit on the margins, looking in), and those which both receive many incoming links *and* frequently link to others (sites which, in network analysis terms, have the greatest centrality in the network, and act as its key connecting hubs). Additionally, networks may also exhibit more or less pronounced clustering patterns, where sites in specific clusters are highly interlinked with one another, but not with sites in rival clusters. In their study of political blogs during the 2004 US election campaign, for example, Adamic and Glance (2005) found that bloggers on either side of US politics interlinked strongly with likeminded authors, but much more rarely connected to their political foes.

Such analysis is of interest well beyond the relatively narrow field of explicitly political blogging, however: it can provide information on the structuration of the blogosphere on a much larger scale, as well as on the internal structures of its smaller subsets that are based around shared interests or identities. It is important to note in this context, however, that conventional Web crawlers tend to be somewhat limited in their capabilities, on several major points: first, they tend not to be able to distinguish between blogs and other types of Websites, and will need to be trained to ignore links which may be commonly present in the sites analysed, but are irrelevant for the study itself (for example to Google search functionality, or to blog platform providers like Wordpress or Blogger). An untreated corpus of links generated by a crawler may therefore include a range of links that need to be removed manually, depending on the specific aims of the network analysis.

Second, crawlers usually also fail to distinguish between different types of links found on the same blog. A typical blog page will contain headers, footers, and sidebars in addition to the blog post itself (and any comments which may have been made by readers) – however, the hyperlinks present in each of these sections fulfil vastly different discursive roles. Links in headers, footers, and sidebars may mainly serve a purely navigational role, allowing readers to access different sections of the blog site, and should usually be excluded from any analysis altogether; some sidebar links (for example in blogrolls) may be relevant, but serve a very

different role from links in the blog post itself: where blogroll links are affiliational (indicating the blogger's longer-term regard for or interest in another site), blog *post* links are much more directly discursive (relating immediately to the topic of the blog post itself, and relevant perhaps only to that post). Indeed, some bloggers might indicate their affiliation with other blogs through the blogroll, but may only very rarely link to actual posts made by those blogs. Any analysis which treats such affiliational and discursive link types as equivalent therefore runs the risk of finding a large number of false positives: showing tight clustering between sites which are on one another's blogrolls, but hardly ever 'speak' to one another in their posts. (Finally, of course, any links provided by commenters on a blog post constitute yet another category – here, in fact, the bloggers themselves are not even the authors of these links, so that the links indicate neither a blogger's affiliation nor their discursive interest in the linked site.) More advanced crawling and capturing technology which reliably distinguishes between these different link types and dismisses those links which do not contribute to the questions at hand is required, therefore, and is being developed by a number of projects, even if it is not yet widely available.

Third, the crawling process itself also provides no opportunity for researchers to set temporal limits on the range of blog posts which are to be included in the corpus, even though this would be of significant interest in many contexts. As we have described it above, crawlers take the URL for a given Webpage, identify the links on that page, and then repeat the process (over multiple iterations) with any of the new pages identified by those links; even if all of the seed URLs are blog posts from a specific period (say, the last week), then, if one or more of those posts link to substantially older posts, the corpus of links produced by the crawler may point to a network of posts which spreads over several months or even years. By contrast, however, researchers may well be interested to study only the interlinkages between blogs during the current week or month – for example to identify which blogs are key opinion leaders on the topics that currently excite the community, and how such opinion leadership may shift over time. Further development of research tools and methodologies will need to address such limitations, and may lead to a move from crawling-based methods towards approaches which capture new blog posts from across a large collection of blogs *as they are made*, and which are thus able to capture the timestamp of those posts; this, then, enables researchers to select for analysis from a larger corpus of data only those posts which were

made within a set timeframe (we describe such a system in Kirchhoff, *et al*., 2009 and Bruns, *et al*., 2008b).

By using other tools than (or in addition to) Web crawlers, it also becomes possible to study more than simply the patterns of interlinkage between the blogs, and to examine the blog content more directly, and on a large scale. If blog posts can be captured (that is, accessed and stored in a local database for further study), this enables the use of tools which provide automatic textual analysis. Tools which enable this are usually a form of Web content scraper – software that takes a list of URLs and captures the content of those Web pages (usually as plain text or HTML). Here, too, it is important to ensure that the scraper distinguishes between the discursive textual content of blog posts themselves and the simply functional or phatic texts found in headers, footers, and sidebars, as well as from comments made by readers – in almost all research scenarios, only the texts of the blog posts themselves will be relevant for the analysis. Further, in using scrapers, researchers must also consider the legal and ethical implications of their actions, of course (which we address in more detail below): for example, whether capturing such content is permissible under relevant 'fair use' or 'fair dealing' provisions in applicable copyright laws, and whether (even though the material captured is openly available online) posts should be de-identified or authors' identities otherwise concealed.

If acceptable approaches can be found, then the generation of a body of textual content – in addition to the information about interlinkage patterns which may be generated by scrapers or similar tools – opens further opportunities for research. Semi-automated textual analysis tools can extract ranked lists of the most frequently used keywords within a specific blog, across a group of blogs, or during a specific timeframe; this provides information on bloggers' core topics (and can be used to examine whether bloggers belonging to different clusters in a link network have different topical emphases) as well as on shifts in overall interest over time (for example in response to external stimuli, such as news or current events). Textual analysis softwares such as *Leximancer* also include the functionality to assess patterns of co-occurrence between individual keywords: this enables researchers to identify clusters of key terms which are commonly used in close proximity, and thus to develop a better picture of topical preoccupations within a single blog (or a group of blogs, or during a set time period)

than a mere ranking of keywords by frequency can provide (*cf.* Bruns, *et al*., 2008a). It may even offer first insights into how specific themes are framed – co-occurrence of the names of specific persons or organisations with emotionally loaded terms ('clever', 'dishonest', 'sexy', 'strong', …) may point to how they are generally perceived by the bloggers whose posts are being studied.

Combining Quantitative and Qualitative Approaches

Such examples also make clear that a merely computational, quantitative approach to blog research is rarely sufficient. Link network mapping and textual keyword analysis, even where they utilise highly sophisticated tools for capturing data at large scale, only provide overall approximations of bloggers' and blog users' activities, and should never be regarded as entirely conclusive in their own right. First, while clustering tendencies in link networks and amongst keyword terms may appear to paint a convincing picture of affiliations and associations, it is necessary at the very least to verify through qualitative spot-checking how such close connection between nodes in the network is to be interpreted. In link networks, for example, frequent and reciprocated linking between two blogs may be the sign of close affiliation – of two bloggers in a cordial and meaningful conversation –, but it could also indicate a bitter, drawn-out war of words. Similarly, the co-occurrence of a political leader's name and key terms such as 'clever' and 'smart' in the textual data could be a sign of high levels of approval – but then, such terms could also have been used in a highly ironic way and mean exactly the opposite. An accurate picture will only emerge if researchers examine at least some of the captured material *in situ* to verify what interpretation is correct.

Second, and more fundamentally, the use of automated tools at least for part of the analysis – while inherently necessary to process the potentially very large datasets which the computational approach can generate – also raises the threat that researchers may treat these tools as black box technologies whose inner workings need not be understood as long as they produce outcomes that 'look right'. Indeed, to significant extent the computational turn could also be described as a shift towards data visualisation – but any even casual glance at visualisation research clearly shows that there is a very wide variety of approaches to turning raw data into graphs. Far from constituting a neutral step in the analytical process, choices made during the data visualisation stage may have significant impacts on how the research

outcomes are interpreted, and may well lead researchers to form conclusions about their objects of study that are not supported by a closer look at the source data themselves. When generating graphical maps of link networks or keyword co-occurrences, for example, researchers must question on what mathematical and network-theoretical basis individual nodes in the network are placed at specific distances from one another, or grouped into particular clusters. This is also a call to form interdisciplinary research teams, of course, combining skills in areas such as cultural studies, communication studies, sociology, and network mapping.

Finally, it is also crucial to note that the data gathered by the methodologies which we have described – however sophisticated the research tools may be – remain only approximations of actual activity by bloggers and their readers. While blog posts are clearly intended as more or less public statements of the blogger's views, we should not fall into the trap of understanding these individuals to be constituted of no more than their stated opinions: their blog content merely represents the – or, more precisely, *one* public face of the author, who may privately or in other public contexts hold some very different opinions. Similarly, in analysing the network of links between blogs what we are actually interested in is a sense of the likely traffic of readers which may follow these links, which in turn provides us with an idea of the respective level of influence of specific bloggers and their points of view over the wider community (of writers *and* readers) populating the blogosphere. Short of gaining access to the often highly prized (and expensively priced) traffic data gathered by online market analysis companies such as Alexa or Hitwise, researchers are able to use linkage patterns as a reasonably accurate indication of likely traffic: both because higher levels of incoming links to a site from other blogs make it more likely that blog readers will find the site, and because more inlinks also increase the site's overall ranking in search engines like *Google*, the assumption that inlinks correlate with traffic is acceptable. Beyond this, we may further expect that the higher visibility that results from a greater number of visitors also provides sites with a better ability to influence and act as opinion leaders for their readers; this necessarily is another assumption only, but one which cannot be tested with any better accuracy unless researchers are prepared to take an excursion into the realms of cognitive science, where the question of the media's effects on audiences remain hotly contested.

Implications of the Computational Turn

Influenced by differing disciplinary emphases and research agendas, the emerging field of online research which bases its work on this computational augmentation (or for some, supplantation) of more conventional social science research methodologies has been described as 'Web Science' (Hendler, *et al*., 2008). This paradigm shift is paralleled by similar moves in the 'Digital Humanities', where the availability of huge, digitised data sets is transforming not only the specific methods, but is reframing the methodologies of particular sub-disciplines as a whole (Manovich, 2007; Moretti, 2005). Manovich (2005) specifies the term 'Cultural Analytics' in order to describe the ways in which computer-assisted approaches transform not only the methods, but in some ways even the object of study of humanities research. While in the present chapter we lack the space to chart in detail the range of epistemological and methodological approaches associated with the 'computational turn', researchers seeking to explore the opportunities inherent in this field would do well to understand the varying configurations of existing discipline knowledge which may be brought to bear on their work; they need to do so with the understanding, too, that methods and methodologies as well as disciplinary framings still remain highly mobile in this new area of research activity.

What is already becoming evident, however, is that whatever the field may be called in the future, one area which needs to be addressed with some urgency is that of research ethics. The computational research tools which are applied in this area are powerful and should not be used without due consideration. Already, over the past few years we have seen major controversies over the public release of major datasets and/or large-scale gathering of personal data – from search engine queries (van Wel and Royakker, 2004) *Facebook* data (Zimmer, 2010a) and *Twitter* updates (Zimmer, 2010b) – using computational tools similar to those we have described for blog research here.

Because of the power of these new research tools and the scale at which they can be employed, these new controversies dramatically exacerbate existing difficulties in applying standard research ethics protocols to online human research. Ethical issues arise at every stage of online research that deals with user-created content and communication. For computer-assisted approaches to blog research, they concern two main interconnected

problems, each of which has implications for how we might understand various choices made in the research process to be ethical. First, are blogs 'publications', or something more akin to personal communication? Second, should individual bloggers therefore be considered to be authors, or treated as research 'subjects'?

To illustrate how this dilemma plays out, computer-assisted blog research may involve gathering, aggregating and analysing the actual content of a large number (hundreds, or even thousands) of blogs. Depending on the form the eventual analysis will take, a decision may need to be made about whether this content can then be quoted (perhaps to illustrate findings), and if so, whether or not the content should be attributed to its author or anonymised. In order to answer these questions, the researcher must address issues of both privacy and of authorial agency (for example, in regard to copyright) with regard to the aggregation and later public availability of the data. This is far from straightforward, however, and has proven contentious even in small-scale qualitative research projects (see an extended discussion in Bruckman, 2002). One point of view would hold that blogs are simply publications, and therefore their contents can be quoted and repurposed within the limits of copyright law; indeed, such reuse of blog material may confer additional attention on the creative work of the blog's author or authors. But another point of view, one that is sensitive to the diversity in actual practices and discourses of bloggers themselves (van Dijck, 2004) might be that while some blogs, particularly those concerned with politics and journalism, are clearly 'public' in nature, others might function as personal media produced with a limited imagined audience in mind, so that further dissemination of blog content in a different context from that originally intended may infringe on the privacy or agency of their authors. It is impossible to determine the answers to these questions in advance, as blogging is culturally complex and diverse, and even in the case of individual blogs the answers may change over time and as a result of other highly specific circumstances. Further, in the context of a very large-scale project relying on automated data-gathering and analysis, the standard solution to ethical dilemmas – seeking and obtaining informed consent from each research participant – is usually impractical. Because of this complexity, standard social science research protocols are rarely adequate to deal with these issues, so it is essential that aspiring blog researchers familiarise themselves thoroughly with the scholarly debates and current thinking on best practice in internet research ethics, in order to take a well-grounded position within them when mounting a case to their ethical review board (or IRB). There are

several authoritative online research handbooks that synthesise these unfolding debates (see, for example Buchanan, 2004; Burnett, *et al*., 2010). Additionally, the Association of Internet Researchers working party on research ethics provides a useful guide to reasonably current areas of consensus and ongoing deliberation among the Internet research community (Ess, *et al*., 2002).

Another problem with the standard ethical review process as it operates in most universities is that, particularly for large-scale, exploratory projects in a field of research that is only now emerging, it is not always possible at the planning stage to anticipate and address all the ethical issues – let alone solutions to them – that will arise in the course of the project – for one thing, the object of study is itself continually in flux. Researchers employing the kinds of large-scale computer-assisted methods we have described here will not only be directly negotiating, but also helping to shape complex ethical issues with each methodological and consequent technological choice they make. Hence, it is important that a reflexive and open approach; that is both informed by and responsive to the debates around internet research ethics be built into the methodology as it unfolds throughout the research process.

Suggested Further Reading

Walker Rettberg, Jill. (2008) *Blogging*. Cambridge: Polity.

One of the first scholars to research blogs and blogging, Walker provides an in-depth introduction to the history of the form, and outlines current themes, issues, and research possibilities.

Rogers, Richard. (2009) *The End of the Virtual: Digital Methods*. Inaugural lecture, University of Amsterdam, 8 May 2009. Available at: http://www.govcom.org/publications/full_list/oratie_Rogers_2009_preprint.pdf (accessed 15 Dec. 2010)

Rogers, the founder of Govcom.org, the independent organisation which makes available the IssueCrawler research tool, describes the computational turn in Internet research methods and outlines its implications.

Baym, Nancy K. (2010) *Personal Connections in the Digital Age*. Cambridge: Polity.

Situating social and mobile media in the wider media ecology of which they are a part, Baym demonstrates the growing importance of non-mass media forms for everyday life,

relationships, and communication. She shows clearly why further research in this field is necessary, and important.

Shadbolt, Nigel, and Tim Berners-Lee. (2008) 'Web science emerges.' *Scientific American* Oct. 2008, pp. 32-37.
Co-authored by Tim Berners-Lee, the inventor of the World Wide Web, this brief article outlines the ideas behind the 'Web Science' initiative for quantitative, data-driven research into uses of the Web. It also contains useful further references.

Questions for Further Investigation

- What analytical frameworks exist for interrogating and evaluating the data which may now be gathered from blogs and other social media platforms in large volumes?
- What differences or similarities exist between specific (topical, demographic, or otherwise defined) segments within the overall blogosphere, and/or between different blogospheres as distinguished by ethnicity, language, geography, or nationality?
- What are the dynamics of blogging: what seasonal patterns does blogging activity exhibit, and how does prominence in the blogosphere wax and wane?
- How do blogs interrelate with other mass or social media spaces: how does information travel across the wider media ecology, and how is blogging affected by developments in more recent social media platforms?
- What are the research ethics of capturing, analysing, and otherwise engaging with content which, while publicly available, might only have been intended for a small personal audience?

References

Adamic, L. and Glance, N. (2005) 'The political blogosphere and the 2004 U.S. election: Divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery 2005*'. Available at: http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf (accessed 19 Jan. 2009)

Berry, D. M. (Ed.) (2010, forthcoming) *The Computational Turn: The Digital Humanities and New Technology*. London: Palgrave Macmillan.

Bruckman, A. (2002) 'Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet', *Ethics and Information Technology, 4*, (3), pp. 217-231.

Bruns, A. and Jacobs, J. (Eds.) (2006) *Uses of Blogs*. New York, N.Y: Peter Lang.

Bruns, A., Wilson, J., Saunders, B., Kirchhoff, L. and Nicolai, T. (2008a) 'Australia's Political Blogosphere in the Aftermath of 2007 Federal Election'. Paper presented at the AoIR 2008 conference, Copenhagen, 18 Oct. 2008. Available at: http://snurb.info/files/aoir2008/Australia%27s%20Political%20Blogosphere%20in%20the%20Aftermath%20of%20the%202007%20Federal%20Election%20(AoIR%202008).pdf (accessed 7th April 2010).

Bruns, A., Wilson, J., Saunders, B., Highfield, T., Kirchhoff, L. and Nicolai, T. (2008b) 'Locating the Australian Blogosphere: Towards a New Research Methodology'. Paper presented at the ISEA 2008 conference, Singapore, 25 July - 3 August 2008. Available at: http://snurb.info/files/Locating%20the%20Australian%20Blogosphere%20(final%20-%20long).pdf (accessed 7th April 2010).

Buchanan, E. (Ed.) (2004) *Readings in virtual research ethics: issues and controversies*. London: Information Science Publishing.

Burnett, R., Consalvo, M. and Ess, C. (2010) *The Handbook of Internet Studies*. London: Wiley-Blackwell.

Burgess, J. (2006) 'Blogging to learn, learning to blog', in Bruns, A. and Jacobs, J. (Eds.) *Uses of Blogs*. New York, N.Y: Peter Lang. Pp.105-114.

Ess, C. and AoIR ethics working group (2002) 'Ethics Guide'. Available at: http://aoir.org/documents/ethics-guide/ (accessed 27th May 2010).

Gurak, L. J., Antonijevic, S., Johnson, L., Ratliff, C. and Reyman, J. (Eds.) (2004) *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Available at: http://blog.lib.umn.edu/blogosphere/ (accessed 27th May 2010).

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. and Weitzner, D. (2008) 'Web Science: An Interdisciplinary Approach to Understanding the Web', *Communications of the ACM, 51*, 7.

Kirchhoff, L., Nicolai, T., Bruns, A. and Highfield, T. (2009) 'Monitoring the Australian Blogosphere through the 2007 Australian Federal Election'. Paper presented at ANZCA 2009, Brisbane, 8 July 2009. Available at: http://snurb.info/files/talks2009/Monitoring%20the%20Australian%20Blogosphere%20through%20the%202007%20Australian%20Federal%20Election.pdf (accessed 7th April 2010).

Manovich, L. (2007) 'Cultural analytics'. Available at: http://www.manovich.net/cultural_analytics.pdf (accessed 24th April 2010).

Moretti, F. (2005) *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Russell, A. and Echchaibi, N. (Eds.) (2009) *International Blogging: Identity, Politics, and Networked Publics*. New York, N.Y: Peter Lang.

Schmidt, J. (2006) *Weblogs: Eine kommunikationssoziologische Studie*. Konstanz: UVK.

Van Dijck, J. (2004) 'Composing the Self: Of Diaries and Lifelogs', *Fibreculture Journal*, *3*. Available at: http://www.journal.fibreculture.org/issue3/issue3_vandijck.html

Van Wel, L. and Royakker, L. (2004) 'Ethical Issues in Web Data Mining', *Ethics and Information Technology*, *6*, pp. 129-140.

Walker Rettberg, J. (2008) *Blogging*. Malden, MA: Polity.

Zimmer, M. (2010a) '"But the data is already public": on the ethics of research in Facebook', *Ethics and Information Technology*, *4*, (3), pp. 205-216.

Zimmer, M. (2010b) 'Is it Ethical to Harvest Public Twitter Accounts Without Consent?' Available at: http://michaelzimmer.org/2010/02/12/is-it-ethical-to-harvest-public-twitter-accounts-without-consent/