Big Data Analysis

Prof. Axel Bruns
Digital Media Research Centre
Queensland University of Technology
Brisbane, Australia
a.bruns@qut.edu.au – @snurb_dot_info

Abstract

‘Big data’ is a current buzzword in the media and communication industries, and in the disciplines that study them, much as ‘Web 2.0’ was in the 1990s. ‘Big data analytics’ is rapidly emerging as a field of research and development, drawing on increasingly rich, increasingly widely available datasets on facets of contemporary life ranging from climate change through economic performance to social media activity. Journalism is caught up in the surge towards ‘big data’ in two ways: as a driver of innovation, through the development of new journalistic specializations currently operating under names such as data journalism or computational journalism; and as an object of analysis, with data on the performance of journalists and news organizations (especially on the public response to their work through social media) being used to justify decisions to increase or decrease staff and funding.

This chapter explores the uses of ‘big data’ in the latter sense, as tools in journalism research, and identifies the threats and opportunities inherent in such developments. It highlights major advances in the field, and shows how the tracking of ‘big data’ on the public resonance of journalistic work provides new evidence about the actual role of journalism in the wider public sphere, beyond normative dogma. At the same time, the chapter also highlights the potential threats which emerge from a strongly quantitative, data-driven turn in journalism research: for example, from the pursuit of strong popularity metrics at the expense of journalistic quality.

Keywords

big data, journalism, social media, quantification, audience reception, user engagement

Bio

Axel Bruns is an Australian Research Council Future Fellow and Professor in the Digital Media Research Centre at Queensland University of Technology in Brisbane, Australia. He is the author of *Blogs, Wikipedia, Second Life and Beyond: From Production to Produsage* (2008) and *Gatewatching: Collaborative Online News Production* (2005), and a co-editor *of Twitter and Society* (2014), *A Companion to New Media Dynamics* (2012), and *Uses of Blogs* (2006). His research examines the uses of social media in political communication, crisis communication, and other contexts, and he is leading the development of new research methods for large-scale social media analytics. His research Website is at http://snurb.info/, and he tweets at @snurb_dot_info.

Introduction: ‘Big Data’ in and on Journalism

The concept of ‘big data’ on virtually all aspects of human endeavour has become a driving force in scholarly research and industry practices across a wide range of fields: *inter alia*, ‘big data’ are being used to inform investment decisions and stock market trades; to support political campaigning and decision-making; to trace the influence of climate change on weather patterns; and to forecast and track the spread of epidemics through the global population. This trend towards the comprehensive quantification both of natural phenomena, of human behaviours, and of complex technological systems has emerged in earnest since the start of the second decade of the new millennium, and is leading to the establishment of entirely new professional roles – from data analysts and even data scientists to the ‘data journalist’: the journalist/researcher who specializes in working with 'big data' sources rather than human informants as the basis of new journalistic stories.

Such trends are substantially informed and even driven by the increased availability of sensors and sensor-like objects that generate detailed and continuous data on a wide range of subjects; these data sources range from sensors measuring the properties of the physical world to computational subroutines logging activities in online spaces. In both cases, significant improvements in the availability and affordability of data storage as well as in the processing power and measurement sensitivity of such sensor objects have made it possible to generate increasingly detailed and up-to-date data on a wide range of measurable properties. Further, the emergence of a field of data processing and statistical analysis which is now often referred to as 'data science' has contributed substantially to the increased combination and correlation of individual sources of data into a greater whole, and thus into 'big data' proper.

Especially where applications of big data analytics focus on the study of human activities and behaviours, data drawn from online sources have come to play an especially important role. The widespread everyday use of Internet-based communication tools at least in developed nations means that data on user activities now represent a substantial portion of total public communication in such countries; as a result, Rogers argues, it becomes possible for Internet studies and allied disciplines no longer merely to study 'the Internet' as a communications space in itself, but in fact to study 'culture and society *with the Internet*' (2009: 29).

Although there are important limits to this broader research agenda, which we will discuss later in this chapter, the potential for Internet studies to pursue such aims is enhanced further not just by the availability of rich data on user activities online, but also by significant advances in the development of powerful and innovative research methods that are able to process and analyze such data. Such methodological developments have been described by Berry (2011) as a 'computational turn' in media, communication, journalism, and social science research: they draw crucially on computational data analytics methods developed in the computer sciences, mathematics, and statistics, as well as on novel computational modelling and forecasting techniques, but they do so in order to apply these methods to long-standing and well-established questions in the humanities and social sciences. This trend, then, also forms an important basis for the emergence of what has been called the 'digital humanities' as a broader scholarly endeavour (cf. Arthur and Bode, 2014).

More narrowly, scholars such as Lev Manovich or John Hartley envisage the development of new research endeavours which they describe variously as 'cultural analytics' or 'cultural science' (Manovich, 2007; Hartley, 2009). These draw on the well-established conceptual and methodological frameworks of cultural studies – and, by extension, of fields such as media, communication, and journalism studies – but add to this the computational, quantitatively focussed, conventionally 'scientific' approaches emerging through the computational turn in digital humanities research methods, in order to provide cultural studies and related disciplines with a more rigorous and more comprehensive evidence base.

Such more thoroughly computationally informed social science research methods clearly have their applications in both journalism practice and journalism studies, too. In the first place, big data are immediately valuable *in* journalism, as the emergence of computational or data journalism as a distinct journalistic practice demonstrates: a direct engagement with and interrogation of increasingly detailed and powerful data sources provides journalists with rich first-hand information that can be used to test the public statements of the stakeholders in a specific debate, and to separate political spin from underlying reality (see the chapter by Lewis in the present volume for a more comprehensive discussion of that trend). But in addition to such uses in journalistic practice, big data *on* journalism also offer important new insights for the news industry and journalism studies alike, by providing new evidence on trends in journalism production and reception at an unprecedented level of detail. In particular, the observation and quantification of Web- and social media-based user engagement with the news generates a number of new metrics that advance well beyond conventional television ratings and print circulation figures, measuring not the broad *distribution* of journalistic content but the specific *uses* made of it, and responses to it. They measure, in short, the agency of news users in much greater detail than ever before, click by click and tweet by tweet.

For better or for worse (and we will examine the threats as well as the opportunities inherent in such big data on journalism later in this chapter), such data enable a quantification of journalistic practice beyond the counting of mere column inches and circulation figures, or of their digital equivalents. Drawing on a combination of internal and external data sources that describe the processes of journalism production and the patterns of audience reception (with particular emphasis, at the current stage of the lifecycle of the journalism industry, on reception by online and social media audiences), practitioners as well as scholars are able to investigate the performance, impact, and relevance of journalists and news organizations at levels of resolution ranging from the individual story through the positioning of specific mastheads to long-term trends in the news industry. In particular, such analysis enables journalism researchers to reveal and benchmark some of the institutional emphases and biases across various news organizations, and to uncover the extent to which editorial decisions or story placements may be driven or influenced by the organizations' own analysis of online audience uses and responses.

But the effective use of such journalism analytics also depends on the quality of the underlying data, the appropriateness of the computational methods used to process the data, the skills of the analyst, and the ability of scholars and industry decision-makers to combine data analytics results with other information sources – a face-value acceptance of journalism analytics as the sole source of knowledge on the contemporary journalism industry is likely to substantially misrepresent the real picture. In what follows, we discuss first the types and potential uses of big data on the production and reception of journalistic content. We then explore the opportunities and threats inherent in an embrace of such data as an important source of intelligence on journalistic practices and the positioning of the news industry in wider public debates, for both journalism practitioners and journalism scholars. Finally, we outline the necessary next steps in pursuing productive and effective, but also considered and critical uses of big data on journalism.

Big Data on the Production of Journalism

Traditionally, at least for researchers without direct access to the newsroom itself, the production of journalistic content has been difficult to measure and quantify, especially on an industry-wide basis. The shift towards an online-first publication of news articles and other journalistic content over the past decades – a shift which is essentially complete by now – has made such measurements considerably easier, however, and there are now a number of readily available means for identifying and tracking the publication of news on a global basis.

RSS and Social Media Feeds

One of the earliest mechanisms for doing so was the use of RSS feeds. Rich Site Summary (RSS) documents are available from virtually all mainstream news Websites, as well as from citizen journalism sites, news blogs, and many other alternative news sources: they do away with the end-user-oriented layout and formatting of news articles on the Web pages of a news site, and present only the core information (usually consisting of article titles, publication dates and other authoring information, a permanent article URL, and an abbreviated article summary) about recently published news articles, in reverse chronological order and in an immediately machine-readable format. Such RSS feeds are intended in the first place for users of news reading software such as the (now discontinued) *Google Reader* or current market leader *Feedly*, where the data contained in subscribed RSS feeds are combined and formatted for more effective use.

But beyond such end-user applications, the RSS feeds of news sites can also be captured and processed for journalism research purposes, where they become up-to-date pointers to the new content published through these sites. RSS scraping software can be used to add the new articles being advertised through such feeds to a continuing database of news articles, which is then available for further processing and analysis that may identify, for example, the daily patterns of news production and publication, or (with additional processing) the key themes and topics addressed in headlines and article synopses. A further extension of this approach would follow the link to the URL that is included in the RSS feed, and capture the full article text for processing as well (cf. Bruns et al., 2008, which describes this approach in detail for capturing blog posts).

This approach may be extended further by also capturing the news updates posted by official news organization accounts or staff journalists to leading social media platforms such as *Facebook* and *Twitter*, and again analysing the new articles promoted through such updates. Standard social media content tracking tools (discussed in more detail below) can capture the tweets posted through the official *Twitter* account of a @guardian or @nytimes, for example, or the posts to the official *Facebook* pages of these publications, and extract from these any links to new articles on the news organization's Website; subsequently, the full publication details and article texts for these stories may be captured from the site itself. Especially in comparison across different social media platforms and with the RSS feed itself, this may reveal different strategies for selecting which articles are promoted through the different social media platforms, for example.

Web Scraping and Article Databases

While RSS as well as social media feeds might provide useful core details on the content production and publishing activities of news organizations, then, direct data gathering from the news sites themselves also emerges as an important approach. Capturing Web content (also known as Web scraping) generates a momentary snapshot of the target Web page as it appeared to the scraper – and by extension, to an ordinary user – at the time of capture; through such scraping, it becomes possible both to gather the full text of a news article, which is not normally included in the RSS feed, as well as a range of important ancillary information which may similarly not be available from other, programmatic data sources. Such information includes the placement and positioning of a given news story on the news site's entry page; the references to further related articles that may be appended to the central story; and also the reader comments that may be published below the story itself. Indeed, repeated scraping of the same page can reveal the dynamics in such additional details: for example, changes in how the story is advertised on the news organization's Website, or the unfolding user discussion that follows the story.

Given the need to separate irrelevant ancillary content (such as advertising, masthead headers and footers, or lists of other popular stories on the same site) from the core information contained in scraped Web documents, and in light of disruptions to the scraping process that may be caused by changing page designs (which require the Web scraper to be retrained in separating core from ancillary content), even automated Web scraping can turn out to be a labour-intensive process, however. Especially where scraping approaches are mainly considered simply in order to capture the full text of news articles, a more workable alternative approach is the use of standard news article databases such as LexisNexis or Factiva: these, too, typically contain the full text of the news article as it was published on the news organization's Website (cf. Wallsten, forthcoming; Vincze, 2014). Given a set of article titles and/or publication URLs for a given publication as they may be retrieved from its RSS feed, therefore, it becomes possible to automatically extract the corresponding full text for each article from the database, and to combine these sources into a new dataset for analysis. (The software and tools to do so may need to be developed on a case-by-case basis, however, depending on the specific target sites and data sources available to the researcher.)

The problems of generating reliable data on the production output of given news organizations through processing RSS and social media feeds and/or scraping news sites and databases may be largely avoided, of course, if researchers have direct access to the internal article databases of the news organization itself: such databases are likely to contain the full titles and article texts (possibly even across multiple revisions), as well as related authoring and publishing details. However, few news organizations are likely to make such databases available beyond internal use – although it should be noted that some global news leaders such as *The New York Times* and *The Guardian* are now offering public Application Programming Interfaces that offer some such information (cf. Toledo Bastos, 2014) –, and it is especially unlikely that outside researchers may gain such access from multiple news organizations, enabling comparative studies of news publication activities. For such industry-wide work, the retrospective establishment of a comprehensive dataset on news production outputs from a combination of RSS feeds and other sources is likely to remain the only feasible – if itself complex and labour-intensive – option.

Further Data Processing

Once gathered for one or more news organizations, such data may then also be processed further ahead of detailed analysis. In particular, article headlines, synopses, and body texts may be subjected to a number of advanced computational textual analysis techniques to establish key content patterns. Keyword occurrence and co-occurrence measurements can be used to generate comparatively simple indicators of the central themes of each article, and such indicators may be aggregated to trace the rise and fall of specific terms and themes over time or show their prominence in specific sections of a publication (cf. Vincze, 2014; Touri & Koteyko, 2014). Additionally, more advanced Natural Language Processing (NLP) techniques can be used for a variety of more sophisticated purposes, including the identification of key named entities – politicians, celebrities, organizations, locations, nations – or attempts at quantifying the sentiment of specific articles.

The outcomes of such further processing may then also be correlated with a range of other data points, of course – for example to examine the relative coverage of specific themes or actors across different news publications, the sentiment towards specific issues across different journalistic authors, or the longitudinal dynamics of such aspects over the course of an extended public debate. Again, yet further comparison and correlation with additional external data sources may also be possible here: such sources may include political polling, economic indicators, casualty figures in current armed conflicts, or other 'official' data, as well as the behavioural data provided by tools such as *Google Trends*, which allows the exploration of global and local trends in Google searches since 2004. In comparison, these different data may provide an indication of the alignment or divergence of journalistic emphases and contemporaneous public opinion.

Potential Uses of Big Data on Journalistic Production

A number of potential uses of big data on journalistic production have already been outlined in passing in the preceding discussion. More generally, both for news organizations themselves and for journalism scholars, a first point of interest in analyzing these composite datasets is likely to be in identifying the trends in publication activity for one or more news outlets. A simple volumetric analysis of news outputs may point to key moments of heightened activity; combined with a first thematic review of article titles and contents, the patterns in output volume may also be able to be linked directly to the prevalent themes in public debate at the time. To the extent that individual journalists can be identified as authors, their specific contribution to the journalistic coverage of these themes can also be examined and quantified.

Beyond such basic metrics on journalistic production, however, more sophisticated analysis techniques – in particular, computational textual analysis – can also be used to provide a more detailed indication both of key themes as such, and of their dynamics over time; similar techniques may also shed light on the centrality of specific individual and institutional actors in public debate, and make such positioning comparable across different news outlets. This may reveal, for example, coverage emphases or biases across different news organizations, and enable the interpretation of such patterns as motivated variously by different news value frameworks, by diverging political ideologies, by editorial decisions to assume a distinct role in setting public agendas, or by the particular audience demographics of specific news outlets. Similar analyses may also allow for an assessment of journalists' and news organizations' sourcing practices: they can identify the prominence of specific sources or source types, from political leaders through domain experts to the *vox populi* in the form of direct interviews or citations of social media posts, and thus shed light on the relative prominence and relevance which different news outlets accord to these diverse sources. Working with data drawn from scholarly news article databases, Wallsten (forthcoming) explores the sourcing of views from social media during the 2012 U.S. presidential election campaign, for example.

It should be noted in this context, however, that such analyses should take care not to treat all news articles contained in their datasets as simply equal: coverage in a lead article placed at the top of a news Website, and promoted widely through social and other media channels, is likely to have made a much more substantial impact on the news audience than reporting in a minor story hidden in a thematic subsection. Here, analyses of story placement (as enabled by the scraping of the front pages of news Websites, and the tracking of social media updates by news organizations, for example) provide important additional data which may serve as proxy indicators for the general visibility of a story, and should be included as multipliers in any comprehensive modelling of thematic and other biases (cf. Lee et al., 2014).

The more complex analytical frameworks which are required for any more detailed approximation of such measures also point clearly to the fact that the capture and analysis of big data on journalistic production alone cannot be the final stage of journalism research; rather, it serves as an enabler of a more advanced mixed-methods research agenda that utilizes big quantitative data and combines them with detailed qualitative investigation. Indeed, this is a fundamental point which is often sidelined by the current public and scholarly discussion about the emergence of 'big data' analytics as a major new research framework: 'big data' methods should not be employed to the exclusion of all other, already well-established qualitative and quantitative methodologies; rather, if deployed appropriately they can complement, support, and integrate with other research methods to enhance the overall quality of the research being conducted. 'Big data' analytics, for example, are able to pinpoint particular observable phenomena which should be singled out for further, qualitative study; conversely, smaller-scale qualitative methods are often indispensable in the development of initial research questions and hypotheses which may then be tested at scale by examining much larger datasets. (See the chapter by Kim Schrøder in this volume for a further discussion of this point.)

Big Data on the Reception of Journalism

Fortunately, it is not necessary to rely solely on an interpretation of news organizations' story positioning and promotion activities in assessing the visibility and impact of specific stories, or in examining a news site's overall market position. Big data on journalism now encompass an especially rich range of sources of data describing the *reception* of journalistic content, in addition to the sources on journalistic *production* which we have already encountered. Such data enhance, extend, and complement pre-existing reception data (including circulation and ratings figures for print and broadcast journalism), especially by providing rich, detailed, and real-time insights into the consumption and use of news content in online and social media.

In this context, it should be noted that the pre-existing reception data for conventional media are not without their own problems. The limitations of broadcast ratings (in the present case, especially for news and current affairs programming) are already well-established: they are extrapolated from often relatively small, demographically representative sample households, and rarely take into account the quality of attention paid to broadcasts (was the TV news on as a background to breakfast or dinner, or did it command its viewers' attention); additionally, they also continue to struggle at taking into account time-shifted and on-demand viewing (Bourbon & Méadel, 2014). Similarly, newspaper circulation figures may variously capture print and distribution runs, or actual subscriptions and purchases, and again cannot provide any insight into how and to what the extent the paper is read after purchase (do readers engage with the news content cover to cover, or do they pay attention only to the politics, sports, or even job advertisement sections). By contrast, the data which can be generated – both internally, by news organizations themselves, and externally, by researchers – for online news reception practices present a considerably more precise perspective on the distribution of different kinds of audience attention; further, in light of the continuing shift towards online news consumption as the dominant mode of access (following especially the increasingly precarious drop in newspaper readership in many markets, and a slower decline in broadcast news audiences in many markets; Christensen, 2013; Pew Research Journalism Project, 2014) makes such indicators especially relevant to the study of journalism audience practices.

Site Access and Activity Data

The online equivalent of circulation and ratings figures for news publications is provided by data describing the volume of Web access attempts to the servers on which news sites are hosted. Such data are available in the first place only to the operators of these servers, and news organizations are already paying increasingly close attention to their performance on these indicators. However, such server data are rarely available to outside researchers, and therefore especially do not allow for an industry-wide benchmarking of news organizations' market positioning. For such purposes, alternative metrics are provided (usually on a commercial basis) by a number of online monitoring services which play a role comparable to that of television ratings agencies: companies such as Experian Marketing Services gather general anonymized information on Internet users' Web browsing practices, and are able to extrapolate from such data a very detailed and demographically representative picture of what Websites are visited by users in specific countries or geographic regions. The largely automated, ISP-level and opt-in panel nature of such data gathering enables them to gather such data on a much larger scale than was possible for television ratings agencies, however: compared to the 3,500 homes included in Australian TV ratings agency OzTAM's data (OzTAM, 2011), for example, Experian Hitwise Australia's online trends data are drawn from some 1.5 million Australians.

Such internally and externally generated data on site accesses are not limited to simple volumetrics alone, however. In addition, they commonly also provide an indication of the upstream and downstream destinations of Web users (that is, the sites they visited before and after the news site itself), of the interactions with the news site (pages visited, time spent), and potentially also of any contributions made in the form of on-site comments, *Facebook* likes, or tweets sharing specific news articles. Especially where users are required to log in to the news site, or where the site is using browser cookies to track individual users, such interactions may also be used to build up a detailed longitudinal profile of user interests across a number of individual visits; this may be done both through the use of internal server analytics tools or by using industry-standard add-ons such as *Google Analytics* (but in both cases, the data generated are unlikely to be available to external researchers or for industry-wide benchmarking).

Social Media Engagement Data

Although such site access and activity data can constitute extremely detailed and highly valuable sources of information on user activity patterns, then, they are rarely available (or at least affordable) for research uses: news organizations tend to treat their internal site statistics and *Google Analytics* data as commercial-in-confidence, and external agencies' products are similarly targetted mainly at commercial users, and often priced beyond the reach of publicly funded projects. A second class of audience engagement data are more readily available, however: these provide insights into how the users of mainstream social media platforms are using and sharing the articles published by news outlets throughout their personal networks.

Such research approaches find their predecessors in earlier studies of linkage and citation patterns in the blogosphere, which tracked the content of a population of known blogs (often focussing on political topics) and identified any hyperlinks in their blog posts. Network analyses of these hyperlink connections were used, for example, to study the relative information sourcing behaviours of blogger populations of different political persuasions (Adamic & Glance, 2005; Park & Thelwall, 2008), or to examine the shift in sourcing practices over time and in response to specific current public debates (Highfield, 2011). However, the relatively disorganized, decentralized structure of the blogosphere – comprised of a large collection of individual Websites as well as blogs hosted on a range of blog platforms – and the widely divergent publishing formats supported by such sites made a truly comprehensive analysis of such patterns virtually impossible; such studies could generally provide only a glimpse of linkage patterns for their specific sample of blogs, therefore, and were unable to make any more comprehensive observations for national or global blogospheres as such. This has changed with the emergence of *Facebook* and *Twitter* as centralized platforms for social media engagement: here, it is possible at least in principle to identify all tweets and all posts that link to a given news site or article, or address a specific topic. The differing affordances of these platforms, and different rules on data access and terms of service, introduce a number of limitations to such possibilities, however.

Facebook

For *Facebook*, the current global market leader in social media, it is generally impossible to identify comprehensively how its users are engaging with specific news services: for the majority of users, access to their *Facebook* activities is available only to approved *Facebook* 'friends' or 'friends of friends', and we must assume that many users who have made their profiles globally public without such restrictions have done so by accident, out of confusion about *Facebook*'s frequently changing privacy controls. Under these circumstances, any data about user engagement (through likes or shares) with news sites that is drawn from currently globally public *Facebook* profiles alone are likely to present a very skewed picture, and should be dismissed as unrepresentative. Similarly, it would be unethical for individual researchers to use their personal *Facebook* credentials to authenticate a *Facebook* data gathering tool and conduct research on the news engagement of their friends without warning them first, and even to do so with the explicit approval of one's friends would again generate an unrepresentative dataset.

What is possible for *Facebook* is to focus solely on the public engagement of users with the official pages of news organizations. Such engagement, even by accounts whose personal activities on *Facebook* are protected by the relevant privacy settings, is public, and it can be assumed that users who do participate on these organizational pages are aware of this fact; this limits (but does not entirely eradicate) any concerns about the ethical acceptability of this research approach. (The development of appropriate ethical guidelines for researchers already has a long history in Internet research; for more details, see especially the work of the Association of Internet Researchers Ethics Committee at ethics.aoir.org.) Proceeding in this way means that researchers are able to gather a number of useful data points about the engagement of *Facebook* users with the news organization, and about the news outlet's own social media activities on the platform: it becomes possible to track the publication of each new post by the page operators, as well as the public response in the form of likes, shares, and comments attached to the post, and to gather some basic information about the overall structure of the *Facebook* audience. A number of tools for gathering data in such a way exist now, including the stand-alone *Facepager* application (Keyling & Jünger, 2013) and the *FacePy* framework for Python programmers (Gorset, 2014).

Twitter

The situation for *Twitter* is somewhat different, due to the different structure and affordances of both the *Twitter* platform itself, and of the *Twitter* Application Programming Interface (API) through which user activity data may be accessed (Bruns & Stieglitz, 2014). Here, the vast majority of user accounts, and their tweets, are globally public and accessible even to non-registered visitors to the *Twitter* site; further, the global *Twitter* 'firehose' of all user tweets is searchable and may be accessed through site and API, at least in principle. This flat and open structure of the *Twitter* network and the user activities that take place within it has also contributed significantly to the predominant role *Twitter* now plays in disseminating and discussing breaking news, even in comparison to its considerably larger rival *Facebook* (Dewan & Kumaraguru, 2014): *Twitter* hashtags, in particular, have played an important role as a gathering point for potentially global *ad hoc* publics around specific issues and events (Bruns & Burgess, 2011).

However, access to the full, unfiltered firehose of all tweets is not generally available to scholarly or commercial researchers, though such access may be purchased at significant cost from third-party data resellers such as Gnip and DataSift (cf. Bruns & Burgess, forthcoming) and may eventually be provided under certain conditions by the U.S. Library of Congress, which was gifted a full and continuing archive of all tweets by Twitter, Inc. in 2010 but has yet to determine whether and how this archive may be made public (Raymond, 2010). But access to *Twitter* activity streams for specific hashtags, keywords, and other search terms is generally available both through the open, free API and through commercial resellers, although the former is limited to providing no more than one per cent of the total current firehose volume, and may therefore be incomplete for hashtags and keywords relating to very significant global events. Further, the API also provides access to underlying user profile and network data which may be valuable in assessing the comparative visibility and impact of different users' engagement activities. A number of stand-alone open-source research tools have established themselves as virtual standards for scholarly research into *Twitter*, including *yourTwapperkeeper* (2012) and *DMI-TCAT* (Digital Methods Initiative, 2014) – both of which require researchers to have access to a Web server for installation. (See Gaffney & Puschmann, 2014, for a more detailed discussion of *Twitter* research tools.)

For our present purposes, such tools may be used to generate a number of key metrics that – similar to the *Facebook* metrics outlined above – describe user engagement with the organizational presences and content of news outlets. In the first place, for any given institutional *Twitter* account it becomes possible to track its own promotional activities by capturing all of its tweets, and to similarly capture any tweets that @reply to or retweet its messages. Especially also across a number of competing accounts tracked using this approach, this enables an assessment of the relative *Twitter* audience response to the account's activities, pinpointing for example which type of tweet (or which type of news article linked to in the tweet) receives the greatest number of @replies or retweets. Similar research can also be conducted around the *Twitter* presences of individual journalists, of course, as Hermida et al.'s study of the role of NPR journalist Andy Carvin during the 2011 Arab Spring demonstrates (2014).

But beyond tracking activity around official accounts themselves, by using the domain names of specific news outlets as search terms researchers are also able to capture those tweets which independently of tweets from those official accounts share links to the stories published on the news site. For example, Bruns et al. (2013) have used this approach to create the Australian *Twitter* News Index (ATNIX), a long-term longitudinal study which has by now generated more than two years of data on link sharing practices around the 35 leading Australian news and opinion sites (fig. 1). On a day-to-day basis, this research approach allows for the identification of currently important themes and topics in Australian news coverage; over time, it enables the assessment both of changing topical interests amongst audience members, and of the gradual evolution of the practice of link sharing itself.

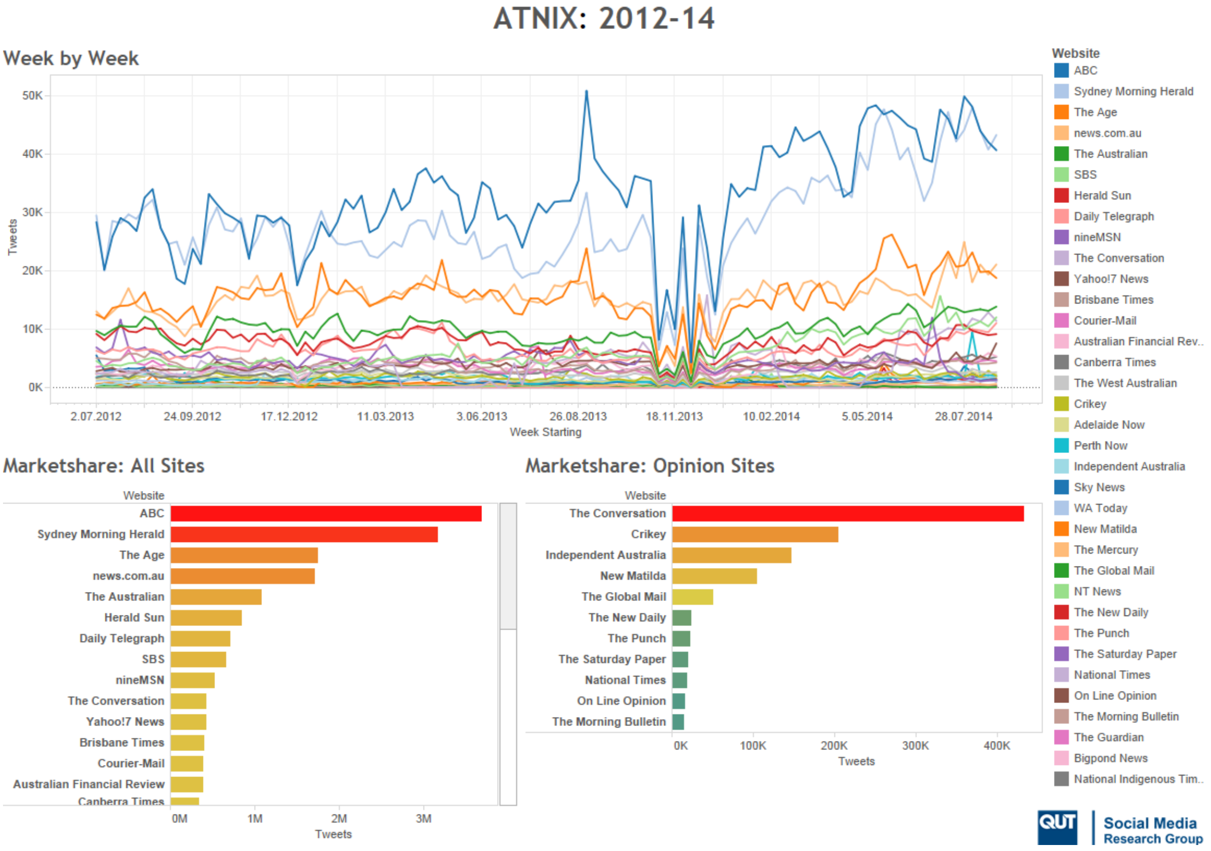


Fig. 1: Sharing of links to Australian news sites, July 2012 to August 2014

Further, the comparatively open structure of *Twitter* also provides a greater range of additional underlying data which may be utilized to assess the likely impact of such individual acts of user engagement. For both *Twitter* and *Facebook*, a number of basic user metrics are available, but only on *Twitter* is it also possible to build up – if slowly, due to API access limitations – a more comprehensive perspective of the overall network of follower connections and of an individual user's positioning within it. Both in assessing the impact of an individual user's actions in engaging with a news brand, and in determining the overall footprint of a news outlet's followers across the Twittersphere, such underlying data are invaluable, as we will discuss in more detail below.

Potential Uses of Big Data on Journalistic Reception

In combination, the large datasets on the reception of journalistic content both in general, and in particular in the spaces of social media, constitute an unprecedentedly detailed source of insight into how Internet users engage with the news – and it is evident that such online access now increasingly constitutes a first point of engagement with the news, ahead of broadcast or print news. Overall, such data shed new light on the total volume of user attention as well as on the comparative prominence of different news outlets and story themes within this emerging attention economy for the journalistic content. Over time, such data may also be used to trace the relative rise and fall in such attention, of course – and may thus also come to influence the future content strategies of news organizations themselves.

Such metrics differ in important ways from conventional ratings and circulation figures, as noted above. Online and social media access generally operates on a 'pull' rather than 'push' basis, where content is deliberately *accessed* by users, rather than broadly *distributed* by publishers, and so the user activities measured by site access and social media metrics have a more deliberate quality than audience metrics for other media. The quantifiable metrics for social media platforms, in particular, truly represent user *engagement* rather than mere readership: here, what is identifiable is whether users choose to like, share, comment on, @reply to, or retweet a news item, while in fact there is no immediately available indicator on whether they have actually *read* the original news article they are engaging with. A framework which fully integrates such engagement metrics with more conventional ratings and circulation figures has yet to be developed by audience research scholars, but a more sophisticated typology of the different forms of news use and engagement is emerging (see esp. Costera Meijer & Groot Kormelink, 2014).

Even in the absence of such a grand unified theory of audiences, these new online metrics provide invaluable new approaches to assessing the positioning and authority of individual news brands, evaluating the performance of their content, and benchmarking such indicators against their competitors in a transparent and scientifically rigorous fashion. For news organizations themselves, the aim of such benchmarking may include the assessment of return on investment for specific activities and initiatives (did major exclusives generate significant readership; do the brand's own social media activities impact on link sharing or site visits); additionally, such research methods may also be used to test the repercussions of specific new initiatives (relating variously to the style of stories, the placement of articles on the site's front page, or the use of search engine optimization techniques and targetted advertising in promoting specific stories or the entire site).

For industry and scholarly researchers alike, further interests may include the identification of specific target audiences for particular content types, forms, and formats, by using the demographic audience breakdowns provided by access data services such as Experian Hitwise; the highlighting of individual highly influential users who serve to amplify the brand's social media presence through disseminating its updates to their own network of followers, by engaging in further network analysis of underlying social media networks; or even the pinpointing of potential expert sources for future news coverage, by identifying those social media respondents who provide consistently useful comments on news articles either on-site or through social media channels.

For journalism scholars, finally, an analysis of the data sources outlined here also offers a unique opportunity to examine the processes of public debate overall, or within specific issue publics (Habermas, 2006; Dahlgren, 2009), within the present-day public sphere or at least those sections of it which operate through public communication in online media channels. Using the approaches outlined here, it is possible not only to examine the activities of the mainstream media (and of the voices commonly enabled to speak within mainstream media coverage), as most previous studies of the public sphere have done, but also to study on a much more comprehensive level the general public's responses to and engagement with such mainstream media content, as well as the interactions between these two levels that may ensue as a result.

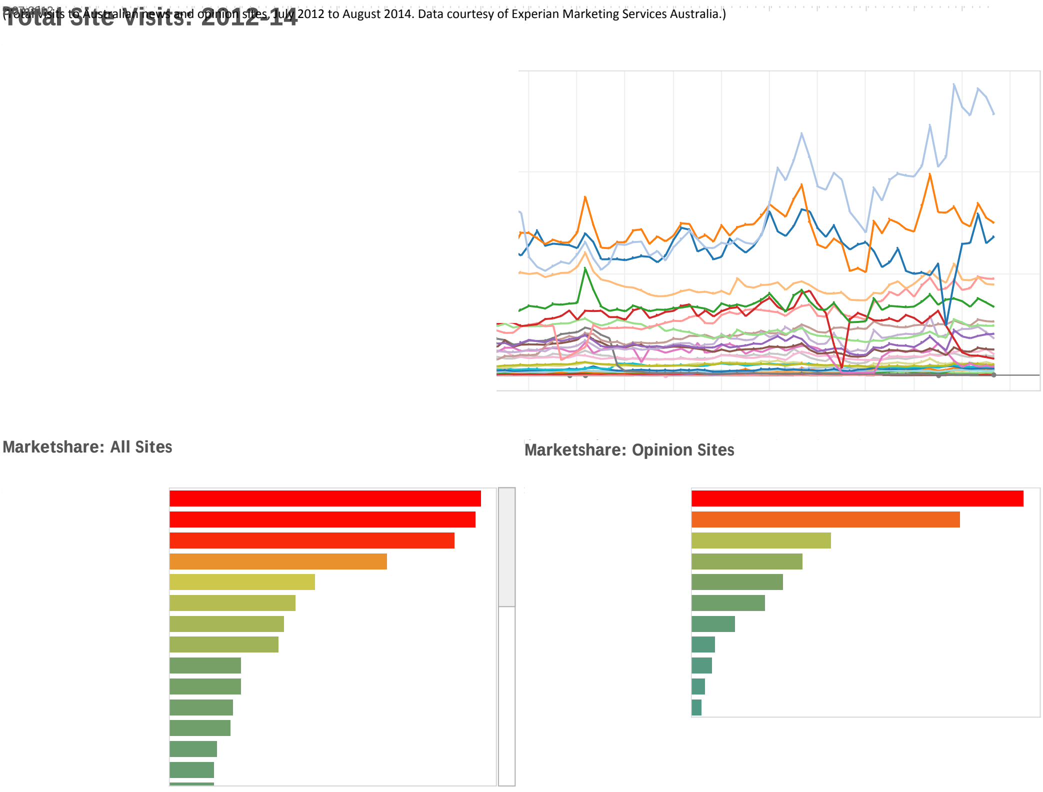
Opportunities and Threats

The big data sources on journalistic production and reception outlined here offer a range of opportunities for both industry practitioners and journalism researchers, then. Drawing on such news sources on audience behaviours, and combining them with their own data on production activities, the news industry is in a position to develop a much more detailed, comprehensive, and evidence-based perspective of its audience's interests and activities, and thus to better tailor and position its news content. At a time of considerable financial strain, this should assist especially in making well-grounded decisions on both funding cuts and new investment.

However, at the same time such a data-driven approach to staffing and funding decisions may not necessarily result in a better journalistic product, even if it generates a more popular news service. There is a danger here that a focus of business decision-making on what might be called 'big data logic' to the exclusion of all else could create a very one-sided type of news organisation. Most centrally, news managers whose decisions are solely driven by return on investment as measured by data on article clicks and shares may find it difficult to justify long-term investment in potentially loss-making activities such as in-depth investigative journalism, even though a news organization's ability to engage in such complex journalistic endeavours may affect its public authority to a considerable degree. There is a danger that such strongly data-driven approaches to determining the journalistic and branding objectives of a news organization would result in a wholesale shift towards populist, attention-grabbing headlines and content which is designed simply to attract short-term readership (and thus to create advertising impressions and generate revenue), even if it fails to build – or even undermines – brand authority. This is an approach which at present is essentially synonymous with the *Buzzfeed* news brand, whose content is inherently designed to 'go viral' on social media (with titles such as 'There's Another 'Game Of Thrones' Theory And It Changes Everything') and which closely tracks the performance of its stories by using methods such as those outlined here. However, it should also be noted in this context that *Buzzfeed* has more recently stated its intentions to use such populist content as a means of drawing in audiences for more sophisticated long-form journalism (Stelter, 2011). The success of this strategy remains to be evaluated – and indeed, the 'datafication' of news organisations' decision-making processes has now become an important new area in journalism research, as the impact of non-traditional news models like *Buzzfeed* on the rest of the industry is being evaluated both quantitatively (through data-driven research into news brands' changing popularity) and qualitatively (by observing or surveying news professionals' attitudes towards these new competitors and their working methods). Of course, beyond mere observation some journalism researchers will also work directly with established and emerging news brands to help formulate their operational strategies in a changing media environment.

For journalism researchers, then, the data sources outlined here constitute important comprehensive and largely independent sources on the production and reception of journalistic content, and support a range of innovative new research agendas. For both aspects of the journalistic process, they enable both the detailed and essentially real-time tracing of the development of a story, as well as the long-term tracking of news articles' themes and news organizations' market positioning. Bruns & Sauter (forthcoming) document the dissemination of a single breaking news item across an increasingly international network of social media users over the course of a few hours, for example, while figs. 1 and 2 document long-term changes in audience attention to differing Australian news sources, measured both by tracking overall access patterns to leading news sites and by tracing the dissemination of links to such sites via *Twitter*.

Where such approaches focus on news production, they are able to engage in what amounts to a reverse engineering of news organizations' coverage agendas and institutional biases by documenting the presence or absence of specific news themes and actors and benchmarking these measurements against their competitors; taking into account the positioning of articles on the news sites' front pages as well as any evidence of search engine optimization strategies in headlines and content, the emphasis placed on such elements by individual outlets as well as the audience response to such initiatives may also be examined. (Research into such questions is not entirely new, but the use of such big data sources as we have encountered them here enables a considerably more comprehensive approach than has generally been possible previously.) By contrast, where research approaches focus on news reception, they are able to quantify the overall popularity of news organizations amongst users by using general site access data (see for example fig. 2 for a year-long study of the relative use of leading Australian news sites, based on Experian Hitwise data); in combination with further social media data, they can also investigate the role and impact of social media link sharing activities on such access patterns.



(Total visits to Australian news and opinion sites, July 2012 to August 2014. Data courtesy of Experian Marketing Services Australia.)

Fig. 2: Total visits to Australian news and opinion sites, July 2012 to August 2014

Overall, these data sources enable researchers to study news organization and audience behaviours ‘in the wild’, outside of controlled laboratory experiments: by drawing on large and detailed datasets, established without affecting the processes they describe, that cover both the content publication practices of news outlets and the access and engagement activities of audiences, journalism studies is for the first time able to investigate news production and reception processes at scale, beyond (but importantly also in useful combination with) smaller-scale case studies, surveys, and interviews. This more systemic perspective on news processes, then, also offers important new perspectives on long-established theories including opinion leadership, the two-step flow, the spiral of silence, or even the public sphere as such: for the first time, it offers strong and large-scale empirical evidence that may be used to examine to what extent such theories still hold or may need to be adjusted to suit the present-day media ecology.

However, progress towards these goals is hampered by a number of important obstacles. First, because of their comparative novelty there is a profound lack of well-established methods for working with these new sources of big data on journalistic practices, and even of comprehensive documentation of the methods employed by the leading research teams utilizing such datasets. Further, while we have focussed here especially on the fruitful combination of a number of these diverse data sources in pursuit of greater research questions, access to such large data remains limited and piecemeal for many researchers, and the combination and integration of diverse datasets is still in its infancy. Especially for datasets which were not created by the researchers themselves, significant black box problems also persist: without commercial operators such as Gnip and DataSift providing detailed documentation on how they gathered and processed their data, there are clear limits to the reliability and usability of these sources.

Worse yet, the relative novelty and allure of such methods obscures the fact that many journalism researchers (as well as researchers in the related fields of media, communication, and Internet studies) lack the methodological training and research expertise to use big data effectively or even correctly. The computational turn in the humanities and social sciences has barely commenced, and many scholars seeking to work with big data on journalism continue to rely substantially on the help of computer scientists, statisticians, and other extradisciplinary colleagues in formulating their methodological and conceptual research frameworks. Such interdisciplinary collaboration can be highly fruitful, of course, and team-based research approaches are generally advisable in dealing with big data, but it remains incumbent on journalism scholars to develop their own methodological skills both in order to collaborate more effectively with these colleagues, and to ensure that their research strategies are appropriate to the project at hand.

Such *caveats* echo the general warnings about 'big data' which have been expressed most articulately by boyd and Crawford (2012). Like theirs, these notes of caution are not intended to dismiss the idea of using big data on journalistic practices altogether, of course; rather, they seek to engender an open and honest discussion about the opportunities and limitations for journalism studies that are inherent in such new datasets and methods. Chief amongst these is perhaps the almost inevitable focus which big data on journalism place on Internet-based modes of journalistic production and reception; while, as we have argued above, online engagement is now often the first form of engagement with journalistic content, an overemphasis on such online modes to the exclusion of all other modes of production and reception necessarily introduces its own biases. There remain significant questions over the extent that researchers can indeed use the Internet as a lens through which they may observe society as a whole – or at least over the amount of distortion that such a lens introduces. In utilizing Internet-centric datasets on journalistic production and reception, we must therefore always also ask what practices are not included in such datasets.

Big Data on Journalism: Where to from Here?

Finally, then, this overview of current opportunities and threats in the use of big data on journalism must necessarily conclude that significant issues and limitations must still be addressed before such analytical methods, and the datasets they build on, can become an everyday part of the journalism researcher's toolkit. There are, very obviously, great opportunities in using big data to further this field of research, but these opportunities will not be able to be fully realized without substantial further methodological and conceptual development. This chapter should therefore also be seen as a call to arms: we must work furiously to develop, test, and document our transdisciplinary skills, methods, approaches, and frameworks for the use of big data in journalism studies, and engage in a frank and open debate about the limits of such approaches – not in order to dismiss them altogether and defend established journalism research practices from this new disruption, but to determine where they may make a useful contribution to the existing methodological toolkit.

Most of all, the use of big data in journalism studies must be more than mere number-crunching. Big data research approaches are wasted if they only serve to provide simplistic measures of volume and size (of news production, of audience engagement); they must advance beyond these metrics to also examine the impact and importance of journalistic and audience practices both for individual news stories, news outlets, and news audiences, and for public debate, the public sphere, and society as a whole. This more sophisticated and comprehensive perspective also ensures that big journalistic data is not used (or abused) simply to justify cost-cutting exercises or drive an increasingly populist repositioning of news brands and their content. Such more complex and in-depth analyses, it should be noted, are also likely to rely on more than mere quantitative data processing: they are set to draw on mixed-methods approaches that utilize both quantitative, big data approaches *and* qualitative, in-depth exploration. Used in this way, then, big data on journalism may also be used to empower journalists and their audiences, rather than merely providing the tools for news organizations to generate better performance indicators.

Acknowledgments

Research for this chapter was supported by the Australian Research Council through my ARC Future Fellowship grant *Understanding Intermedia Information Flows in the Australian Online Public Sphere*, with data on total visits to Australian news and opinion sites courtesy of Experian Hitwise Australia. My thanks also go to my QUT colleague Darryl Woodford for advice on current data gathering facilities for *Facebook*.

References

Adamic, Lada, and Glance, Natalie (2005, 4 Mar.) 'The political blogosphere and the 2004 U.S. election: divided they blog', *Blogpulse* (http://nielsen-online.com/downloads/us/buzz/wp_PoliticalBlogosphere_Glance_2004.pdf).

Arthur, Paul Longley, and Bode, Katherine (eds) (2014) *Advancing Digital Humanities: Research, Methods, Theories*. Houndmills, Basingstoke: Palgrave Macmillan.

Berry, David (2011) 'The computational turn: Thinking about the digital humanities', *Culture Machine*, 12: 1-22 (http://www.culturemachine.net/index.php/cm/article/ view/440/470).

Bourbon, Jérôme, and Méadel, Cécile (eds) (2014) *Television Audiences across the World: Deconstructing the Ratings Machine*. Houndmills, Basingstoke: Palgrave Macmillan.

boyd, danah, and Crawford, Kate (2012) 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication and Society*, 15(5): 662-679. DOI:10.1080/1369118X.2012.678878.

Bruns, Axel, and Burgess, Jean (forthcoming) 'Methodological innovation in precarious spaces: The case of Twitter', in Helene Snee, Christine Hine, Yvette Morey, Steven D. Roberts, and Hayley Watson (eds), *Digital Methods for Social Sciences: An Interdisciplinary Guide to Research Innovation*. Basingstoke: Palgrave Macmillan.

Bruns, Axel, and Burgess, Jean (2011) 'The use of Twitter hashtags in the formation of *ad hoc* publics', paper presented at the European Consortium for Political Research conference, Reykjavík, 25-27 Aug. 2011 (http://eprints.qut.edu.au/46515/).

Bruns, Axel, and Sauter, Theresa (forthcoming) 'Anatomie eines Trending Topics: Methodische Ansätze zur Visualisierung von Retweet-Ketten', in Axel Maireder, Julian Ausserhofer, Christina Schumann, and Monika Taddicken (eds), *Tagungsband Digital Methods*.

Bruns, Axel, and Stieglitz, Stefan (2014) 'Metrics for understanding communication on Twitter', in Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (eds), *Twitter and Society*. New York: Peter Lang. pp. 69-82.

Bruns, Axel, Highfield, Tim, and Harrington, Stephen (2013) 'Sharing the news: Dissemination of links to Australian news sites on Twitter', in Janey Gordon, Paul Rowinski, and Gavin Stewart (eds) *Br(e)aking the News: Journalism, Politics and New Media*. New York: Peter Lang. pp. 181-210.

Bruns, Axel, Wilson, Jason, Saunders, Barry, Highfield, Tim, Kirchhoff, Lars, and Nicolai, Thomas (2008) 'Locating the Australian blogosphere: Towards a new research methodology', paper presented at the ISEA 2008 conference, Singapore, 25 July - 3 Aug. 2008 (http://snurb.info/files/Locating%20the%20Australian%20Blogosphere%20(final%20-%20long).pdf).

Christensen, Nic (2013, 8 Nov.) 'ABCs: Newspapers see more double digit declines', *Mumbrella* (http://mumbrella.com.au/abcs-newspapers-3-188553).

Costera Meijera, Irene, and Groot Kormelink, Tim (2014, 1 Aug.) 'Checking, sharing, clicking and linking: Changing patterns of news use between 2004 and 2014', *Digital Journalism*. DOI:10.1080/21670811.2014.937149.

Dahlgren, Peter (2009) *Media and Political Engagement: Citizens, Communication, and Democracy*. Cambridge: Cambridge UP.

Dewan, Prateek, and Kumaraguru, Ponnurangam (2014) 'It doesn't break just on Twitter: Characterizing Facebook content during real world events', *arXiv*: 1405.4820 [cs.SI] (http://arxiv.org/abs/1405.4820).

Digital Methods Initiative (DMI) (2014, 12 June) *Twitter Capture and Analysis Toolset (DMI-TCAT)* (https://wiki.digitalmethods.net/Dmi/ToolDmiTcat).

Gaffney, Devin, and Puschmann, Cornelius (2014) 'Data collection on Twitter', in Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann (eds) *Twitter and Society*. New York: Peter Lang. pp. 55-68.

Gorset, Johannes (2014) *FacePy* (https://github.com/jgorset/facepy).

Habermas, Jürgen (2006) 'Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research', *Communication Theory*, 16(4): 411-26.

Hartley, John (2009) 'From cultural studies to cultural science', *Cultural Science Journal*, 2(1): 1-16.

Hermida, Alfred, Lewis, Seth C., and Zamith, Rodrigo (2014) 'Sourcing the Arab Spring: A case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian revolutions', *Journal of Computer-Mediated Communication*, 19: 479-499.

Highfield, Tim (2011) *Mapping Intermedia News Flows: Topical Discussions in the Australian and French Political Blogospheres*. PhD thesis. Brisbane: Queensland University of Technology (http://eprints.qut.edu.au/48115/).

Keyling, Till, and Jünger, Jakob (2013) *Facepager: An Application for Generic Data Retrieval through APIs* (https://github.com/strohne/Facepager/).

Lee, Angela M., Lewis, Seth C., and Powers, Matthew (2014) 'Audience clicks and news placement: A study of time-lagged influence in online journalism', *Communication Research* 41(4): 505-530. DOI:10.1177/0093650212467031.

Manovich, Lev (2007) ‘Cultural analytics’ (http://www.manovich.net/cultural_analytics.pdf).

OzTAM (2011) ‘About OzTAM ratings’ (http://www.oztam.com.au/AboutOzTAMRatings.aspx).

Park, Han Woo, and Thelwall, Mike (2008) ‘Developing network indicators for ideological landscapes from the political blogosphere in South Korea’, *Journal of Computer-Mediated Communication*, 10(4): 856-79.

Pew Research Journalism Project (2014, 26 Mar.) ‘Key indicators in media & news’ (http://www.journalism.org/2014/03/26/state-of-the-news-media-2014-key-indicators-in-media-and-news/).

Raymond, Matt (2010, 25 Apr.) ‘How tweet it is!: Library acquires entire Twitter archive’, *Library of Congress Blog* (http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/).

Rogers, Richard (2009) *The End of the Virtual: Digital Methods*. Amsterdam: Vossiuspers UvA (http://www.govcom.org/publications/full_list/oratie_Rogers_2009_preprint.pdf).

Stelter, Brian (2011, 12 Dec.) ‘BuzzFeed adds politico writer’, *The New York Times*: Media Decoder (http://mediadecoder.blogs.nytimes.com/2011/12/12/buzzfeed-adds-politico-writer/).

Toledo Bastos, Marcos (2014) ‘Shares, pins, and tweets: News readership from daily papers to social media’, *Journalism Studies*, 5 Mar. 2014. DOI:10.1080/1461670X.2014.891857.

Touri, Maria, and Koteyko, Nelya (2014) ‘Using corpus linguistic software in the extraction of news frames: Towards a dynamic process of frame analysis in journalistic texts’, *International Journal of Social Research Methodology*, 3 July 2014. DOI:10.1080/13645579.2014.929878

Vincze, Hanna Orsolya (2014) ‘“The Crisis” as a journalistic frame in Romanian news media’, *European Journal of Communication* 29(5): 567-82. DOI:10.1177/0267323114541610.

Wallsten, Kevin (forthcoming) ‘New media in the newsroom: Twitter as a news source during the 2012 campaign’, *Newspaper Research Journal*.

yourTwapperkeeper (2012) (https://github.com/540co/yourTwapperKeeper).