Hunters and collectors: seeking social media content for cultural heritage collections

Paper presented at the VALA2014 17th Biennial Conference, Melbourne.
http://www.vala.org.au/vala2014-proceedings/vala2014-session-7-barwick

Kathryn Barwick
Project Officer, Innovation Project
State Library of New South Wales
kathryn.barwick@sl.nsw.gov.au

Mylee Joseph
Project Leader, Innovation Project
State Library of New South Wales
mylee.joseph@sl.nsw.gov.au

Cecile Paris
Research Leader, Computational Informatics
Commonwealth Scientific and Industrial Research Organisation
cecile.paris@csiro.au

Stephen Wan
Research Scientist, Computational Informatics
Commonwealth Scientific and Industrial Research Organisation
stephen.wan@csiro.au

Abstract:
A novel approach to collecting digital content for heritage collections is being explored and assessed in a trial of Vizie, an innovative social media tool researched and developed by the Commonwealth Scientific and Industrial Research Organisation. Collecting digital content for heritage collections is a priority for research libraries and other cultural institutions. This paper reports on the progress and learnings to date of the ongoing collaboration between the CSIRO and the State Library of New South Wales. The aim of the collaboration is to gather and curate online content centred around significant events and every day life in Australia and New South Wales.

Introduction

Memory institutions, such as state and national libraries, are mandated to collect and preserve documentary heritage for future generations. The collecting role of institutions, including libraries, archives and museums, is expanding rapidly with the increasing volume of digital content created. This content includes objects that are converted into digital forms and also content that is born digital across a variety of formats. Existing approaches to digital collecting, for example PANDORA Australia's Web Archive (Koerbin 2012), target Australian online publications in a highly selective way. These individual selections include some digital ephemera, but there is a strong emphasis on official, published digital content from organisations, institutions, agencies and businesses. A gap in the documentary record exists both globally and in Australia, as existing digital collecting tools generally do not include social media content.

In the early 21st century, social media plays an increasingly important part in Australian public life providing an avenue for public comment, political debate, information sharing, humour and, perhaps most importantly, unfiltered public opinion. This high volume, dynamic data set in the form of tweets, messages, shared images and video, posts, comments and other forms of interaction including curation and crowd sourcing forms part of our shared digital heritage and is of interest to libraries on behalf of the future researchers they seek to serve. The priority to source, sample and archive some of this ‘community created content’ is challenging libraries like the State Library of New South Wales (the Library) to develop digital collecting frameworks and workflows, and to identify tools to support this type of activity. This paper reports on the progress and learnings to date of the ongoing collaboration between the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and the Library. The aim of the collaboration is to gather and curate online content centred around significant events and everyday life in Australia, particularly New South Wales.

Social media and our collective memory

Social media is proving to be a very valuable resource for researchers from a variety of disciplines (Kietzmann, Silvestre, McCarthy & Pitt 2012), offering windows into the attitudes, reactions and beliefs around current events, social and technology trends, memes, natural disasters and professional activities including conference sound bites and quotes. The consideration of the free labour involved in creating social media as a driver in producing value is also of interest to political economists (Wittel 2013), while the role of social media trends to influence political sentiment, supplement polling and engage the community in political debate (Gil de Zúñiga, Jung & Valenzuel 2012) has been explored in detail since the 2008 Obama US presidential campaign made extensive use of social media (Cogburn & Espinoza-Vasquez 2011).

The place of social media in cultural collections

Social media has similarities with other types of material collected by research libraries and other cultural institutions. Not unlike traditional forms of publishing,

social media is published by an individual with intent to reach an audience of followers and engage in a wider conversation. Marwick and Boyd (2011) describe this as an "imagined audience" where the real audience is both networked and participatory, in the model of "many-to-many communication". Social media also has characteristics in common with the public opinion and communication that was previously conveyed in letters and diaries; these artefacts are routinely included in historical research collections. Frequently the historical record available to us does not provide insights into how the community felt about politics of the day, yet where these types of communications are found, we have greater insight into the mood and attitudes of a community. For example, a comparison between the micro-political statements contained in graffiti of the ancient world (Zadorojyni 2011) and the commentary on politics today conducted via social media could be drawn.

From a collection management perspective, social media content is variously regarded as digital ephemera and grey literature (Gelfand & Lin 2011). The imperative to collect a sampling of it as part of documentary heritage has been recognised in a number of library organisations including the National and State Libraries Australasia (NSLA), a collective which published a digital collecting framework (2013) highlighting the need to include social media content in this cultural dataset. Further afield, the British Library's web archiving activities will include the "cultural and intellectual output that appears in digital form, including tweets and blogs" (Sillito 2013), while the Library of Congress has accepted the entire Twitter archive as an addition to their digital collections (Allen 2013). Collecting digital content for heritage collections is a priority for research libraries and other cultural institutions.

With respect to collecting, it is a priority for the Library "to create [and maintain] a collection that reflects the cultural heritage of New South Wales in both the Australian and international contexts" (State Library of NSW 2013). Personal communications and ephemera are an important part of this heritage as demonstrated by items held in the Library's collection. For example, the survival of several thousand diaries of World War I service men and women, held at institutions including the Australian War Memorial and the Library, gives researchers an insight into the lives of the men and women who participated in the conflict. Ephemera relating to elections, campaigns and political events are very transient, yet provide valuable context for researchers. While the digital equivalents of these communications are recognised as important for collections, the technical challenges remain in identifying, harvesting, archiving and preserving them.

The language of electronic networked publics

Each social media tool has developed an etiquette and style of expression that is unique to the format and channel. The development and use of hashtags on Twitter as an effective means of coordinating a distributed discussion (Bruns and Burgess 2011) is a valuable tool for researchers and also for identifying social media content. Hashtags have spread across other social media tools and can be found in Pinterest, Instagram and Facebook, as well as into interactive use with other forms of media like television (Harrington, Highfield and Bruns 2012). Understanding how language is used in these environments and developing tools to locate and filter content presents a new challenge for identifying, retrieving and collecting information in these spheres. For example, the Twitter hashtags used during the 2012 U.S.

elections were analysed to see how hashtags gain popularity (Lin et al., 2013). In related work, there has been much research recently in attempting to determine the political orientation in users of Twitter (for example, see Cohen and Ruths, 2013, and Wong et al. 2013).

Monitoring and filtering

A number of automated tools are used in the humanities to analyse social media. These can be used by curators to generate an overview of data collected in order to fine-tune the curation process. For example, tools like Latent Semantic Analysis (Deerwester et al., 1990) and Latent Dirichlet Allocation (LDA) (Blei et al. 2003) which find repeated patterns in the data and groups data points based on these patterns. LDA in particular has been gaining popularity as an analysis tool for social scientists (examples below). It discovers a pre-determined number of topics from the data and clusters the documents according to these topics. Social scientists use the discovered topics to obtain a data-driven overview of a large text collection to suggest the salient themes that seem to be represented, at least with respect to the statistical analysis.

Topic modelling approaches like LDA have also been exploited to analyse topics on social media. Ramage et al. (2010) were amongst the first to explore the use of these statistical methods on short texts like Twitter. Ritter et al. (2010) explore methods to adapt LDA to capture the discussion structure of Twitter data. Zhang et al. (2007) use LDA to discover sub-groups in an online community. Bauer et al. (2012) use LDA-related methods to model linguistic patterns with respect to geographical features using Foursquare[i] data. To find topics in historical text, Yang et al. (2011) use LDA to study newspapers from the 19th century to find historical trends in Mexican culture. One variant of LDA, ccLDA tools, creates cross-cultural analyses that provided overviews of the differences in word use between cohorts of online contributors for a particular topic, in this case, tourism discussion participants in the UK, India and Singapore (Paul and Girju, 2009). LDA and ccLDA analyses are available in the CSIRO prototype, but are not currently used in the current data collection phase. In future collection efforts, these tools can be used to identify important terms associated with a theme thereby revealing important query terms that may be used to refine the data collection process.

Collecting digital content for heritage collections

Beyond the technical challenges, collecting institutions encounter a range of other issues. The challenges associated with collecting and archiving social media content as documentary heritage include technical dimensions, workflows, collecting frameworks and the ownership (copyright and IP) of the content itself. Allen (2013) from the Library of Congress remarks that the technology to support "this type of research scholarship access to large data sets lags behind technology for creating and distributing such data". The rapid change in these emerging technologies and the associated terms of service of each social media tool vary, and may also be supplemented by different terms of use for an Application Programming Interface (API) provided to explore and expose content contained within a social media channel. These constraints apply to the ownership of the content, copyright and

reuse or republication, as well as the volume and frequency of obtaining information via an API.
There are also quite different expectations and attitudes to 'ownership' of social media content amongst users (Marshall and Shipman 2011) "most crucially when we investigate how social media is saved or archived; how it is reused; and whether it can be removed or deleted". There is also confusion evident in the comments contributed by readers to the Library of Congress blog posts about the acquisition of the Twitter archive including:

> "1) What rights do I have to the archived tweets if any?
>
> 2) What limitations does the Library of Congress have one [sic] distributing the information if any?
>
> 3) What are the restrictions on private tweets? Can the Library archive them? If archived are there limitations as to whom these tweets are available to? Considering the fact that I intended for them to be private." (SK 2013)

This confusion about the status of social networks as public or private (Boyd 2007), or some kind of hybrid space, persists.

The high volume of content generated by these sources and emerging channels, the speed at which new content emerges and the transient nature of the content make traditional curated models of content selection unworkable. Twitter, for example, carries an average 58 million tweets per day from more than 550,000,000 registered users (Statistic Brain 2013), yet within 6 - 9 days the search function provided by the Twitter Search API (2013) is incomplete and focuses on relevance rather than completeness. Twitter is also an example of social media content that may not be available for public collection, such as posts made on private accounts or other geopolitical constraints.

Some models, such as that used by the NSLA for the PANDORA Archive, are heavily curated, relying on individual identification of web content for archiving according to the selection guidelines developed by each institution involved. The archive prioritises government publications and academic content, and does not explicitly target social media content. The international nature of social media also means that activities like the Library of Congress hosting the Twitter archive (2013) and the British Library whole of domain harvesting are likely to capture content of relevance to NSW heritage collecting as well. In this case no replication would be required if access was available to researchers. The next section summarises the collaboration between the Library and the CSIRO, with respect to Vizie, a collaboration that may go some way towards helping the Library fill the gap in collecting digital content.

A tool for collecting digital content

The aim of the collaboration between the Library and the CSIRO is to gather and curate online content centred around both significant events and everyday life in Australia and New South Wales, using a CSIRO social media monitoring prototype called Vizie. Vizie is designed to collect, analyse and archive social media. It was initially designed to help support government departments in identifying opportunities to improve the delivery of services.

As part of the CSIRO Early Adopters Group trial of Vizie, Library staff have applied Vizie to a new application, testing Vizie for collecting purposes, with a focus on content-based curation. Key features of the Vizie tool include meta-search functionality to gather content from sources such as Google Alerts[ii], Social Mention, Twitter; rule-based relevance filtering, automatic keyword tagging, social media post classification, and the archiving capability for collected posts and monitoring actions. This ongoing collaboration provides a laboratory to test and refine a collecting framework based on the style of language used in online public discussion (via social media environments), workflows for refining search terms (to reduce the non-relevant data retrieved), as well as the continuous development of the features and capability of the tool itself.

There are multiple facets to a framework for collecting social media content for a heritage collection:

[i] designing an adequate set of query terms to represent the event;

[ii] refining query terms to account for ambiguity;

[iii] sampling data from the web with these query terms;

[iv] vetting the returned content for its appropriateness in the collection; and, finally,

[v] the archiving.

Vizie is used in this project to help select content that can be archived with respect to some set of queries (described in a Collection Framework). Vizie facilitates this task by providing:

[i] overviews that highlight commonalities amongst content from different social media platforms, and

[ii] extractive summaries of a post to highlight salient content.

Meta-search functionality is provided in Vizie by allowing a user to enter a query that is then farmed out to the APIs of the various social media platforms like Twitter, Facebook, YouTube, LinkedIn and Instagram. In addition, Vizie also uses the API of the search engine Social Mention to collect data. This is particularly helpful in collecting data in the recent past if there are technical difficulties, either within Vizie or with a specific API. Data collection can be refined using additional constraints on the queries such as imposing an ordering of multiple query terms and using exclusion terms to refine the query.

The Vizie system also follows links in the data where possible to capture linked data. For example, if a social media post like a Tweet mentions the query term but is also part of an ongoing discussion, Vizie will attempt to download the conversational context for that post. Similarly, if a post includes a shortened URL, the link is resolved and that extra web content is included in the Vizie data capture since it provides the context for the social media commentary. Snapshots of the webpage are also taken where possible to archive the content given the transient nature of the web. It should be noted that data archived by Vizie is not available for public dissemination. Data collected and any subsequent analysis of this data is solely for the user of Vizie.

In on-going work, for some APIs, CSIRO is currently researching different methods to use geotagging metadata to limit content to the Australian region. Implementing such filters is challenging because not all users of social media enable the geolocation features. Currently, limiting data to the NSW region is implemented through the judicious use of query terms. Geolocation restrictions may in some cases filter out relevant material, such as tweets about an event that happened or is happening overseas that affects NSW residents, reactions of NSW residents to events happening outside NSW, or overseas discussion of an event in NSW. For this reason, geolocation restrictions have not been enabled for most Library queries, with some exceptions for queries that generate a high number of false positives from overseas sources and where filter words have not been effective in minimising false positives.

Keywords are selected by using a sliding window (currently a week of data) to determine trending words. Vizie also takes into account specific terms that are of interest to the user, which are available to the Vizie analysis system through user-defined word lists. Key phrases are determined using both the keywords and orthographic features, such as capitalisation. Keywords and phrases are then used to cluster and label the data for the overview, and to help select sentences for extractive summaries.

The data collected is automatically tagged with a categorisation system that organises the queries used with the social media platforms. These categories, called “Monitoring Activities”, typically specify the different foci of a data collection exercise. These groupings help divide the data collected from the queries into different subsets, each of which can be analysed and reported on separately. These subsets of data can also be exported for the user to analyse with other social media tools.

As outlined above, Vizie provides the Library with a technical tool to collect social media content, but the Library also needs to create guidelines for its use.

A collecting framework

A curated model of collecting digital content, like PANDORA, is effective with highly targeted selection criteria, search engines for identifying content and a manageable volume of content to examine and archive. Whole of domain harvesting is effective where a domain can be isolated to a geographic area, but many social media tools are hosted on international platforms that do not have these geographic characteristics to assist collecting. In addition, many social media users choose not to disclose their geographic location (Leetaru et al. 2013). The volume of social media content generated also requires a different approach to collecting. Social media content is often “of the moment”, reflecting an individual’s perspective at that moment in time.

Preservation of content posted on social media sites is also subject to the actions and even survival of the social media tool, e.g. Facebook, Twitter, etc, and their decisions in archiving and preserving user-generated content. When dealing with high volume content, for example, popular hashtags on Twitter, it is often difficult to retrieve material that is more than one week old. Consequently, capturing a snapshot

of discussions as they are happening is crucial for the development of a meaningful collection of social media content. Any collecting framework for social media should be robust enough to handle frequent change, and should be in a format that facilitates regular updates.

The Library's existing Collection Development Policy (2013) and list of collecting priorities shaped the collecting framework for social media. The framework also relates to the NSLA Digital Collecting Framework (2013) and the Library's PANDORA selection guidelines (2013). In devising the Library's Social Media Collecting Framework content, experts across the Library were engaged to provide input into the development of the topic areas to be included. Reflecting the ever-changing nature of the material the Library is trying to identify and collect, the framework itself is intended to remain in perpetual beta.

The Collecting Framework sets out aims and assumptions about social media, and the constraints of the operating environment:

- Aims:
    - to identify the types of information to be collected from social media sources in accord with the Collection Development Policy of the State Library of NSW;
    - to capture a data set from a variety of social media sources (eg. Facebook pages, Twitter, blogs, etc.);
    - to reduce the risk of significant personal historical records and social commentary being absent from the Library's collection, and;
    - to capture NSW public sector information distributed on social media channels.
- Assumptions:
    - the audience for the collection will include researchers from a variety of disciplines;
    - not everything can be captured and stored;
    - privacy will be fully investigated before the data sets are made publicly available, but the current imperative is to capture the information while it is still available;
    - some facets of social media information map closely to the other original materials, grey literature and ephemera formats and collecting priorities in the Library.
- Constraints:
    - it is acknowledged that some search strings yield content that is unintended or 'false positive'
    - this is a sample data set, it cannot be comprehensive in nature but it is intended to be balanced and representative of a range of perspectives in issues relating to life in NSW wherever possible
    - the data set is not retrospectively curated
    - Vizie has a quota on queries imposed by APIs (for example, Twitter)

Some of the technical and resourcing challenges encountered throughout the collecting process include:

[i] designing an adequate set of query terms to represent the event, and identifying topics that are trending and the terms used by people talking about the topic:

For example, *#bushfire, #nswfires* and *#sydneyfires* all appeared on 10 September 2013 describing fire activity on the outskirts of Sydney. This requires surveillance of the social media conversation, familiarity with the social media environment and tools for interrogating it, including search engines like Socialmention.com and Topsy.com, and flexibility to change as these tools and the environment change, for example, the closure of Google Reader in 2013. Some query terms can be more easily predicted, including the names of organisations, official Twitter account names and major events. In many cases, multiple queries are required to capture content (e.g. *"State Library" NSW* and *"State Library" New South Wales*). Due to limitations on the number of active queries that can be handled by the Twitter API, a seasonal calendar was developed to manage queries for annual events. These queries are deactivated during the year, and activated only when the event is held. Queries for once-off events are also deactivated. Ongoing conversations can be captured in some of the generic query terms that are left active year-round. For example, *festival sydney* is active year-around and collects discussions about any festivals being held in the Sydney area, while specific queries for festivals such as VIVID and Sydney Writers Festival are active only when the festivals are on.

[ii] refining query terms to account for ambiguity and reducing the false positives captured requires regular checking and vetting of the content captured, and the capacity to turn off queries as many are seasonal topics

For example, *#mardigras* captured highly relevant content during the Sydney Mardi Gras in March 2013, but two weeks later the same hashtag retrieved content about Mardi Gras events in other countries. In advance of the 2014 Sydney Mardi Gras, the event organisers have begun socialising a specific hashtag #sydneymardigras but there is no compunction for members of the public to choose to use this hashtag in preference to one of their own design. The limited number of characters allowed in Twitter (140) often leads to abbreviated hashtags being preferred. Researching the hashtags in use for events is a key activity. As a further example, the global nature of Twitter can often draw in content that is unintended, and refining is an ongoing challenge; for example, Gallipoli is understood in a specific context by most Australians, yet many Italian Twitter users are using the term for describing their beach vacation.

[iii] sampling data from the web with these query terms

Posts collected can be false positives: posts that include the query but are not relevant to the target topic. To alleviate this problem, the Vizie prototype was developed to include a preview function has been added

to the interface. This feature allows queries to be previewed to check for false positives before they are collected. Vizie has also been upgraded to allow the impact of filter words on a query to be previewed. The nature of social media means that false positives can still be collected; however, Vizie's preview function and filter words significantly lessen the number of false positives collected. Query strings are also tested to identify the most effective search terms for retrieving content. Socialmention.com, Topsy.com and Google searches are all used to test queries.

[iv] vetting the returned content for its appropriateness in the collection

Daily monitoring of content collected has been necessary to check for false positives. Where these are detected, there are two options for dealing with them: filter words may be applied to the query that collected them to attempt to filter out these posts from future collecting, or, the query can be deactivated. For example, the hashtag *#heatwave* was used in January 2013 by many Australians during conversations about the hot weather and bushfire threats. By early February, however, content was being collected about the Miami Heat, a professional basketball team based in Miami, USA, as the same hashtag was used by their supporters to encourage their team. As they then had a 27 game winning streak, the volume of posts from their supporters was significant. Several filter words were entered in Vizie, however daily monitoring of the content collected revealed that these were not successful in significantly reducing the false positives collected. This was due to the variety of different terms being used by Miami Heat fans, as the conversation moved so quickly that it was impossible to predict which terms would be used and filter them out in time. Posts from Miami Heat were high volume, significantly higher than any Australian use of *#heatwave* to discuss the weather. At the same time, usage of the hashtag by Australians in relation to Australian topics was in decline. Consequently, the query term was deactivated in March. Strategies for this vetting include frequent monitoring of posts collected (generally multiple times a day), and responding by refining search terms or turning them off but not removing content already collected. The aim is to check for false positives and to minimise future collecting of these sorts of items.

[v] the archiving

In the pilot project, a method of archiving the content is still under investigation. It is acknowledged that the terms of service of the APIs used to retrieve content via Vizie have different restrictions and reuse may have constraints. To further complicate matters, the APIs can change over time (for example, see the notice to Twitter developers 2013) in the content they retrieve and also in the terms of service for their use. Aggregations of social media content like Hashsard.com feature a disclaimer "All tweets captured here have at one time been available on the public record." recognising that account owners can change the status of their account from public to private and users can

also delete posts. The Library of Congress White paper (2013) indicates that the Library is "working to develop a basic level of access that can be implemented while archival access technologies catch up", and cannot yet provide access to researchers. The collecting framework used and the metadata attached to items collected by the State Library via Vizie will also be important to provide context to future researchers using the dataset.

Case study: *#auspol* and *#spill*

Politics is a topic that is hotly debated on social media networks in Australia. The hashtag *#auspol* is used, particularly on Twitter, for discussing and sharing information about Australian politics and political news. Collecting this dataset provides the opportunity to analyse public comment about and reaction to political developments. Hansard, the record of parliamentary proceedings, and media archives can be matched up with social media content to provide an overview of political announcements and developments, media reactions and "spin" on these events, and public reaction, both to the events themselves, and to the events as presented by the media. The hashtag #auspol is in constant use throughout the year and other hashtags are added to the conversation for specific events, including *#spill, #sausagesizzle, #democracysausage* and *#destroyingthejoint*. During the election year, 2013, Twitter Australia reported 250,000 tweets on 21 March 2013 when a spill was mooted by Prime Minister Gillard. Subsequently, 500,000 tweets using the *#spill* hashtag appeared on 26 June 2013 when the change of leadership occurred. In the same way that the political graffiti in Pompeii tells researchers about attitudes to political leaders of the day, the political debate played out on social media networks presents a wide range of perspectives and supplements the official record of Hansard and the perspectives of media outlets.

Conclusion

The Vizie tool is still in the development phase. Even so, it has provided the Library with a focus for digital collecting in the category of social media and the opportunity to explore in detail what type of content can be gathered in this space. Content collected has proved to be a useful snapshot of life in New South Wales, as experienced and shared on social media and for some query terms more broadly in Australia. The social media posts collected cover topics including the 2013 Federal election, bushfires across the summers of 2012-2013 and 2013-2014, discussion of alleged drug use in Australian sport, reactions to current events and festivals, public comment on the Royal Commission into Institutional Responses to Child Sexual Abuse, and much more. The Library is continuing to use Vizie to collect social media content, during the ongoing development of the tool. Next steps for the Library are under consideration.

The experience of being part of a research and development partnership rather than purchasing software off the shelf has also provided many opportunities to discuss how technical challenges could be addressed. This complex collecting challenge draws on the various activities of commercial, government and memory institutions intersecting in the digital collecting and archiving space. The Internet Archive and the Wayback Machine, together with other web archiving activities around the world,

form part of the collective memory of our times. The British Library domain harvesting, PANDORA Australia's Web Archive, Digital NZ and the Library of Congress hosting of the Twitter archive, together with the digital collecting activities of the Library, will provide many data sets for researchers of the future. Other emerging commercial tools, such as the Topsy.com search feature for the Twitter archive, may also provide some access to social media content for researchers. This is a rapidly changing and developing field and there may be other tools emerging in the near future requiring libraries to maintain a flexible and adaptive stance in their digital collecting activities.

References

Allen, E 2013, 'Update on the Twitter Archive at the Library of Congress', *Library of Congress Blog*, web log post, 4 January, viewed 11 September 2013, http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

Blei, D M, Ng, A Y, & Jordan, M I 2003, 'Latent dirichlet allocation', *the Journal of machine Learning research*, vol. 3, pp. 993-1022.

Boyd, D 2007, 'Social network sites: Public, private, or what', *Knowledge Tree* vol. 13, no. 1, pp. 1-7.

Bruns, A & Burgess, J 2011, 'The use of Twitter hashtags in the formation of ad hoc publics', *6th European Consortium for Political Research General Conference*, 25 - 27 August 2011, University of Iceland, Reykjavik. Available at http://eprints.qut.edu.au/46515/

Cohen, R & Ruths, D 2013, 'Political Orientation Inference on Twitter: It's Not Easy!' in *The Proceedings of the International Conference of Webblogs and Social Media*, Boston, USA.

Cogburn, DL & Espinoza-Vasquez, FK 2011, 'From networked nominee to networked nation: Examining the impact of Web 2.0 and social media on political participation and civic engagement in the 2008 Obama campaign', *Journal of Political Marketing*, vol. 10, no. 1-2, pp. 189-213.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W. & Harshman, R. A. 1990. 'Indexing by Latent Semantic Analysis', *Journal of the American Society of Information Science,* 1990*, 41*, 391-407.

Gelfand, J & Lin, A 2013, 'Grey Literature: Format Agnostic Yet Gaining Recognition in Library Collections', *Library Management*, vol. 34, no. 6/7, p. 10.

Gil de Zúñiga, H, Jung, N & Valenzuela, S 2012, 'Social media use for news and individuals' social capital, civic engagement and political participation', *Journal of Computer-Mediated Communication*, vol. 17 no. 3, pp. 319-336.

Goel, V 2013, 'If Google Could Search Twitter, It Would Find Topsy', *New York Times*, web log post, 4 September, viewed 6 September 2013, http://bits.blogs.nytimes.com/2013/09/04/if-google-could-search-twitter-it-would-find-topsy/

Harrington, S, Highfield, T & Bruns, A, 2012. "More than a backchannel: Twitter and television." Ed. José Manuel Noguera. Audience Interactivity and Participation. COST Action ISO906 Transforming Audiences, Transforming Societies, Brussels, Belgium. 13-17.

Hashsard, viewed 12 September 2013, http://hashsard.com/

Kietzmann, JH, Silvestre, BS, McCarthy, IP & Pitt, LF 2012, 'Unpacking the social media phenomenon: towards a research agenda', *Journal Of Public Affairs*, vol. 12, no. 2, pp. 109-119, doi: 10.1002/pa.1412

Koerbin, P 2012, 'PANDORA - past, present, and future National web archiving in Australia', paper presented at the *National Conference on eResources*, 5-7 December, Malaysia, viewed 6 September 2013, http://www.nla.gov.au/openpublish/index.php/nlasp/article/viewFile/2592/3042

Lin, Y, Margolin, D, Keegan, B, Baronchelli, A & Lazer, D 2013 '#Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtags' in *The Proceedings of the International Conference of Webblogs and Social Media*, 8-11 July, Boston, USA.

Leetaru, K, Wang, S, Cao, G, Padmanabhan, A & Shook, E 2013, 'Mapping the global Twitter heartbeat: The geography of Twitter', *First Monday*, vol. 18, no. 2, doi: 10.5210/fm.v18i5.4366

Library of Congress, 2013, *Update on the Twitter Archive At the Library of Congress*, viewed 13 September 2013, http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Marshall, CC & Shipman, FM 2011, 'Social media ownership: using twitter as a window onto current attitudes and beliefs', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 7-11 May, Vancouver, pp. 1081-1090.

Marwick, A E 2011, 'I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience', *New Media & Society*, vol. 13, no. 1, pp. 114-133.

Mishne, G & De Rijke, M 2006, 'Language model mixtures for contextual ad placement in personal blogs', *Advances in Natural Language Processing*, vol. 4139, pp. 435-446.

Mitchell, L, Frank, M R, Harris, K D, Dodds, P S & Danforth, C M 2013, 'The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place', *PLOS ONE*, vol. 8, no. 5, doi:10.1371/journal.pone.0064417

National and State Libraries Australasia (NSLA) 2013, *Digital Collecting Framework*, viewed 21 August 2013, http://www.nsla.org.au/publication/digital-collecting-framework

Ramage, D, Dumais, S & Liebling, D 2010, 'Characterizing Microblogs with Topic Models' in *Proceedings of the Fourth International AAAI Conference on Webblogs and Social Media*, 23-26 May, Washington.

Ritter, A. Cherry, C. & Dolan, B 2010, 'Unsupervised Modeling of Twitter Conversations' in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2-4 June, Association for Computational Linguistics, Los Angeles, pp. 172-180.

Sillito, D 2013,'Libraries to store all UK web content', *BBC News,* 5 April, viewed 5 September 2013, http://www.bbc.co.uk/news/entertainment-arts-22028738

SK 2013, web log comment, 1 February, on Allen, E 2013, 'Update on the Twitter Archive at the Library of Congress', *Library of Congress Blog*, web log post, 4 January, viewed 11 September 2013, http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

State Library of New South Wales 2013, *Collection Development Policy*, viewed 12 September 2013, http://www.sl.nsw.gov.au/about/policies/docs/slnsw_collection_development_policy.pdf

State Library of New South Wales 2013, *PANDORA Selection Guidelines*, viewed 12 September 2013, http://pandora.nla.gov.au/guidelines.html

Twitter Developers blog 2013, *Calendar of API changes*, viewed 16 September 2013, https://dev.twitter.com/calendar

Twitter Developers blog 2013, 'Using the Twitter Search API', *Twitter Developers blog*, web log post, 5 August, viewed 11 September 2013, https://dev.twitter.com/docs/using -search

Wittel, A 2013, 'Digital labor: the internet as playground and factory', *Information, Communication & Society*, doi: 10.1080/1369118X.2013.829512

Wong, F, Tan, C, Sen, S & Chiang, M 2013, 'Quantifying Political Leaning from Tweets and Retweets' in *The Proceedings of the International Conference of Webblogs and Social Media*, 8-11 July, Boston, USA.

Yang, T.-I.; Torget, A. & Mihalcea, R. 2011, 'Topic Modeling on Historical Newspapers' in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, pp. 96-104.

Zadorojyni, A 2011, 'Transcripts or dissent? Political graffiti and elite ideology under the Principate', in: JA Baird & C Taylor (eds.), *Ancient graffiti in context,* Routledge, New York.

Zhang, H, Qiu, B, Giles, C L, Foley, H C & Yen, J 2007, 'An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks' Intelligence and Security Informatics', in *Intelligence and Security Informatics, 2007 IEEE,* 23-24 May, New Brunswick, pp. 200-207.

Endnotes

[i] Foursquare is a mobile application that allows users to post their location at a venue and connect with friends.

[ii] For most of this collaboration, Google Alert content was available as an RSS feed that allowed content to be aggregated and analysed by Vizie. At the time of writing, Google no longer provides this functionality for Google Alerts.