VALA2012 Session 2 Sherratt

university of canberra trove tim sherratt

VALA2012 Session 2 Sherratt

Mining the treasures of Trove: new approaches and new tools

PEER-REVIEWED

VALA2012 CONCURRENT SESSION 2: Discovery
Tuesday 7 February 2012, 11:25 - 11:55
Persistent URL: http://www.vala.org.au/vala2012-proceedings/vala2012-session-2-sherratt

Tim Sherratt

University of Canberra, ACT

Please tag your comments, tweets, and blog posts about this session: **#VALA2012 and #S2TS**

VALA2012-Session-2-Sherratt-Paper (363.11 kB)

View the presentation on the VALA2012 GigTV channel
Tuesday, February 07, 2012, 11:25 PM AUSEDT, 33 Minutes 16 Seconds.

Mining the treasures of Trove: new approaches and new tools

Tim Sherratt
tim@discontents.com.au

Abstract
Recently, the National Library of Australia added the 50 millionth article to its Trove newspapers database. This is an astonishing resource to anyone interested in Australia history and culture. But how do we use it? Historians armed with traditional methods of search and browse now have to 'grapple with abundance', but in doing so, new questions start to arise. How might we track not a person or an event, but an idea? This paper will introduce the possibilities of text-mining and report on some of my experiments in applying these to Trove.

Introduction

> A newspaper is a universal book containing the very essence of literature and science. 'Tis a history of past and passing events... (Anon 1828)

Our view of the print media may not be quite as rosy as that expressed by *The Australian* in 1828, but still most of us would be happy to admit that newspapers provide a 'rough draft of history'.

This now familiar phrase was probably used for the first time in December 1905 by *The State* (Pettinato 2010). 'The newspapers are making morning after morning the rough draft of history', claimed an article on the role of newspapers in education: 'Later, the historian will come, take down the old files, and transform the crude but sincere and accurate annals of editors and reporters into history, into literature' (Anon 1905).

Newspapers were waiting to be mined by historians for the raw material to build their interpretive edifices. However, not everyone agreed. In his address before the American Historical Association in 1908, William Nelson (1908) recalled that George Bancroft, the prominent American statesman and historian, had exclaimed, 'But you cannot write history from newspapers!' Undeterred, Nelson confidently demonstrated how eighteenth-century newspapers could enrich our understanding of revolutionary America. Any historian who ignored this source, Nelson concluded, 'will miss a great and invaluable mass of material'.

Through the changing styles and fashions of history over the past century, newspapers have remained one of the historian's key resources. A study by Dalton & Charnigo (2004) found that 72% of historians surveyed regarded newspapers as important sources for their research. This was backed up by citation analysis that revealed newspapers were the fifth most frequently cited source, after books, manuscripts, journal articles, and government documents.

Helen Tibbo's (2003) 'Primarily History' project focused on historians' use of primary materials. Here newspapers emerged as both the 'most important' and 'most used' sources, with some respondents noting that 'period newspapers were the only source of information that existed on aspects of their research'.

While the importance of newspapers as a source seems to have changed little, the ways in which we access them has changed dramatically. The quote I used at the start of this paper was found in minutes, from the comfort of my home, through a simple keyword search in the Trove newspapers database.

This paper examines what this new ease of access means for historians' use of newspapers. Beyond mere convenience, digital access enables us to frame new questions and seek new answers. Using existing online technologies we can extract and manipulate structured data to build interfaces and visualisations that provide us with useful means of engaging with large cultural resources. This paper presents some examples of tools and approaches based on the Trove newspapers database. What possibilities are emerging and what problems remain?

Rewind / Fast-Forward

As a postgraduate historian I developed a close familiarity with the foibles and frustrations of microfilm readers. My research into the relationship between science and progress drew heavily on newspapers and magazines. I used existing indexes to guide my investigations, and developed a list of events and dates to focus my browsing. But still, newspaper research entailed many days learning the intricacies of spools and lenses, not to mention sore eyes, headaches and the constant battle to get and hold the 'best' machines.

As studies by Allen and Sieczkiewicz (2010) and Jones (2009) show, historians use newspapers in a variety of different ways. Commonly, newspapers are used to obtain more detail about specific events and to situate a study within its broader social context. Historians will often move backwards and forwards between newspapers and other primary sources as they seek to round out a story.

Microfilm access to newspapers constrains this process in a number of ways. Firstly, of course, you need to be where the microfilm is. This places practical limits on the time you can spend searching. Secondly, you either need to work from a list of known dates or be prepared to invest considerable effort in browsing.

Known dates can be extracted from subject indexes, but for Australian newspapers these are fragmentary and rather idiosyncratic. Alternatively you might follow up references obtained from books, articles or archives. In either case, you are limited to events that have already been identified. There is a danger of circularity here — the supposed significance of events can be reinforced by that fact that they are more readily accessible.

Browsing offers the possibility of discovery, but demands a substantial investment in time and patience. As a result of this researchers are likely to focus their efforts on a limited number of titles. Jones (2009) notes that the *New York Times* is over-represented in US research, both because it has a useful index and because it is perceived as a newspaper of record. For similar reasons, the *Sydney Morning Herald* is often the first port of call for Australian historians.

The digitisation of newspapers not only helps us to avoid these sorts of constraints, it has the potential to bring significant changes to historical practice. Ready access offers more than mere convenience. While historians can now conduct newspaper research in their pyjamas while drinking a comforting beverage, the more general point is that newspapers can be integrated into the historian's research processes in flexible and opportunistic ways. Instead of waiting until their next trip to the library, a historian can quickly pursue a hunch or follow a lead. Immediate feedback encourages greater exploration.

Trove already boasts a high proportion of local and regional newspapers. Unhindered by the need to prioritise their browsing, historians are better able to move beyond a narrow, metropolitan focus and pursue their research wherever it may take them. New possibilities emerge for comparing the experience and opinions of the city and the bush.

Perhaps the most revolutionary change is something we now take for granted — keyword searching. Tim Hitchcock (2008) describes how keyword searching can turn the hierarchical descriptive systems of archives on their head, exposing the lives of ordinary people. Newspapers have a less rigid structure, but keyword searching

offers a similar ability to browse from the bottom up, enabling us to read against the grain of contemporary prejudice.

For example, historian Kate Bagnall is interested in documenting intimate relationships between Chinese men and white women in nineteenth-century Australia. Many standard histories will simply deny that such relationships existed, falling back on familiar stereotypes of the Chinese community (Bagnall 2011b). However, such relationships often attracted the attention of a prurient press eager to regale its readership with tales of the hardship and woe it was assumed were the natural outcomes of these unnatural pairings. Simple keyword searches, such as 'chinese AND white', bring these articles to the surface, revealing crucial names and dates (Bagnall 2011b).

Digitisation brings new modes of discovery, but as Ronald Zweig (1998) points out, this can have its drawbacks. Historians used to fossicking for a few relevant articles can be confronted by 'unmanageably large quantities of data':

> It is possible to drink from a stream, but it is impossible to be refreshed by the flood caused when the gates of a dam are opened. (Zweig 1998 p.89)

Zweig argues that advances in digitisation need to be accompanied by developments in informational retrieval. We need new tools to tame the flood.

The challenge of abundance

In a 2008 discussion on 'the promise of digital history', Dan Cohen noted that 'nearly every day we are confronted with a new digital historical resource of almost unimaginable size' (Cohen et al 2008). This was a challenge to the current practices of historians. Could the traditional close reading of a limited number of sources continue to mount a compelling argument without embracing the contextual richness of these ever-burgeoning online offerings? Historians needed new tools, new training and a new way of thinking — 'a methodology for the infinite archive' (Turkel 2006).

This is not simply a matter of building a better search interface. As Cohen (2008) explained, there are a whole range of technologies that could be brought to bear upon these digital storehouses. These include: 'text mining, document and topic clustering, automatic audio and video transcription, and other techniques based on the machine-readable nature of digital materials'. Names can be extracted, places can be mapped, and datasets can be linked. Once in digital form, these resources are open to all manner of manipulation and visualisation. Instead of just finding new things, we can see old things in new ways.

One particularly interesting aspect of these new technologies is that they thrive on data. In 'From Babel to knowledge', Cohen (2006) notes: 'As the size of a collection grows, you can begin to extract information and knowledge from it in ways that are impossible with small collections'. The larger a collection, the easier it is to observe patterns.

But where do you start? Fortunately, historians are not the only ones interested in extracting meanings from large bodies of text. Within the broader field of the Digital Humanities, scholars have been using computers in the analysis of texts for more than half a century (Hockey 2004). In the 1940s, Father Roberto Busa starting working with IBM on the development of a concordance to the works of Thomas Aquinas (Busa 1980). Many have followed in his wake, developing a range of

sophisticated tools and techniques for natural language processing. Much development is also happening outside the academy, with everyone from advertisers to security agencies keen to distil patterns from the masses of textual data available online.

Literary scholars have long been using computers to calculate word frequencies and collocations, or to extract and view keywords in context. These same techniques can readily be applied to historical resources, enabling researchers to quickly construct overviews or time series to guide their investigations. Similarly, developments in named entity recognition allow large quantities of text to be mined for references to people, places and events — the most commonly used access points for historians.

Increasingly, this is not a matter of building new tools, but of extending and integrating existing ones. An interesting example of this is provided by the 'Datamining with Criminal Intent' project, funded by the US Office for Digital Humanities through its 'Digging into Data Challenge' (With Criminal Intent 2011). The raw material for the project was the digitised proceedings of 197,745 criminal trials held at London's central criminal court between 1674 and 1913 — made available through the Old Bailey Online.

As a first stage, a public Application Programming Interface (API) was developed to provide machine-readable access to the Old Bailey proceedings. Instead of developing a custom toolset on top of this API, the project used it to communicate with the text analysis web service VoyeurTools, via Zotero, a research management program (Cohen 2011). As a result, the 'Criminal Intent' project has not only opened the Old Bailey proceedings to sophisticated forms of text analysis, it has fostered developments in Zotero and VoyeurTools that will be of more general use.

Dan Cohen (2008) suggests that 'digital history can be defined as the theory and practice of bringing technology to bear on the abundance we now confront'. Already the Trove newspaper database contains almost 60 million articles with more being added all the time. This is a remarkable collection of individual stories, but it is more. It is also a record of long-term cultural shifts, of changes in the way we think and speak. The challenge is now to embrace the possibilities of digital history and broaden our idea of discovery. There are patterns to be found.

Thinking like a machine

The interface for the Trove newspaper database is, not surprisingly, designed for human beings. Human beings are good at picking up on contextual cues. Amidst a complex page of text and images, we can easily recognise that one string which represents the title of the newspaper in which the current article was published.

A machine, however, would struggle with the same, simple task.

We can help machines. We can tell them how to extract meanings from the long strings of characters they see on a web page. The simplest approach would probably be to give a computer a list of all the newspaper titles in the database and tell it to see if any of them appear on the current page. However, starting with a pre-compiled list seems like cheating, and simply finding text matches offers little understanding of context.

We could build algorithms that look for the types of words and phrasings that normally appear in newspaper titles. We could refine this further by training our

computer with a series of specially marked-up examples. This is an example of natural language processing, or more specifically of named entity recognition. We would be teaching the computer to understand and manipulate the structure of the text itself.

But web pages generally have some sort of structure. They might use <h1> tags to denote a heading, or assign a specific CSS class to style the newspaper title. We can use these structures and the patterns they contain to point machines to specific packets of data, like names or dates, contained within the text. The packets can then be extracted and used. This is known as screen scraping.

The problem with screen scraping is that you need to have a set of recipes for each web page you want to extract data from. And if the underlying HTML changes, then your screen scraper might stop working. It is an inherently fragile technology, but it is a fairly quick and easy way to move from text to structured data — from human-readable to machine-readable.

ScraperWiki (2011) offers a good example of the power and flexibility of screen scraping. The ScraperWiki site gives users a simplified framework within which they can create and test their screen-scraping recipes. Once completed, the recipes and the extracted data are shared with others. Many government data sets, liberated from ordinary web pages, are made available in this way.

Ideally all online resources would expose structured data in machine-readable forms, as APIs and Linked Open Data. Significantly, an API is already under development for Trove. But why should we wait? Scraping is one of the core skills Bill Turkel (2005) recommends for new students: an invaluable part of the toolkit of any 'impatient historian'. Scraping empowers researchers to access and manipulate the data that is important to them. Research questions need not be constrained by institutional priorities (Sherratt 2011g).

So, although Trove is currently designed for human beings, we can do something about it. By poking around in its HTML innards, we can find and expose the newspaper article metadata we need for other forms of machine processing. We can build scrapers and we can experiment.

Harvest time

One of the most empowering aspects of the Digital Humanities is the encouragement to do it yourself. If an existing tool does not quite meet your needs you can either hack it or build a new one. In the case of Zotero (2011), an open-source research manager, users are invited to enhance its functionality by contributing their own specialised screen scrapers. So if Zotero does not yet know how to save metadata from your favourite web resource to your research database, you can write and share a 'translator' that tells Zotero exactly what to do.

I wanted to be able to grab details of newspaper articles from Trove and save them into my research database along with PDF copies of the articles themselves. So I wrote a Zotero translator to do just that. This translator is now part of the main Zotero distribution.

Once saved into Zotero you can of course organise, annotate, tag and share the articles you have collected. You can also take advantage of the built-in timeline visualisation or use a plugin to map any places mentioned in your text. As with the

‘Criminal Intent’ project you can send your references off for text analysis. You can also export in machine-readable formats, and easily import into the web-publishing platform Omeka to create an instant exhibition (Sherratt 2011c).

But while it is easy to revel in the thrill of finding and collecting new articles, the most exciting, and daunting, aspect of the newspapers database is its sheer scale. What do you do when your search returns 10,000 results? 100,000? Even using Zotero you are going to have to do an awful lot of pointing and clicking to save these into a machine-processable form. Do you simply assume the relevance ranking has done its job and concentrate on the first few pages of results? Or do you sample across the set?

To make it easier to explore complete result sets, I created a harvester to download all the article metadata returned by a query in the newspapers database (Sherratt 2011a). The harvester simply loops its way through each page of results, using a screen scraper to extract the details from the pages of individual articles. The metadata is saved as a text file in CSV format so that it can be opened as a spreadsheet, or imported into a database. You are then free to search, filter or analyse the results in any way you see fit.

The harvester also gives you the option of saving the text content of all the articles in a separate zip file. Similarly, you can choose to automatically download and save PDFs of all the articles. Using the harvester can save you a lot of time: just feed it your Trove newspaper query and it will present you with metadata and article contents, all neatly organised and ready to browse at your leisure.

But it lets you do more. By saving the text of the articles into a separate file, the harvester makes it easy for you to analyse their content. To get started, you can simply upload the zip file to VoyeurTools (2011) a powerful, web-based text-analysis environment. Using natural language processing, VoyeurTools can generate an overview of the complete contents of any body of text, including a familiar word frequency cloud. From this workbench, you can drill down to examine words or phrases in detail, or generate a variety of visualisations to explore patterns and relationships within the text.

If your needs are more specialised, you can fire up the text analysis software of your choice and just point it at the zip file. Mallet (2011), for example, enables you to draw out themes from a variety of texts using a process known as topic modelling, while WordSeer (2011) explores the grammatical structures of texts.

The point is that by harvesting the articles in this way you are creating new opportunities to explore, analyse and understand their content.

Big pictures

Harvesting is great if you know what you are looking for. But sometimes you are just not sure. You do not really want to download several thousand articles only to find there is a serious flaw in your query construction. Sometimes what you need is a quick overview, a visual summary of your search that will highlight anomalies and suggest possibilities for refinement.

I decided an easy way of achieving this would be to plot the number of articles returned by a particular query over time. Rather than having to retrieve the details of every article, all I needed was the total number of results for each year. Instead of

spending hours downloading thousands of articles, I could create an overview in minutes (Sherratt 2011b).

Once again my screen scraper extracted the necessary data; all I had to do was repeat the query for every year in the requested span. My script then wrote the data to a JavaScript file which could be easily imported into an html page and graphed using a JavaScript visualisation library.

As well as testing preconceptions and hunting down anomalies, these graphs offer a quick way of comparing the occurrence of words or phrases over time (see Figure 1).

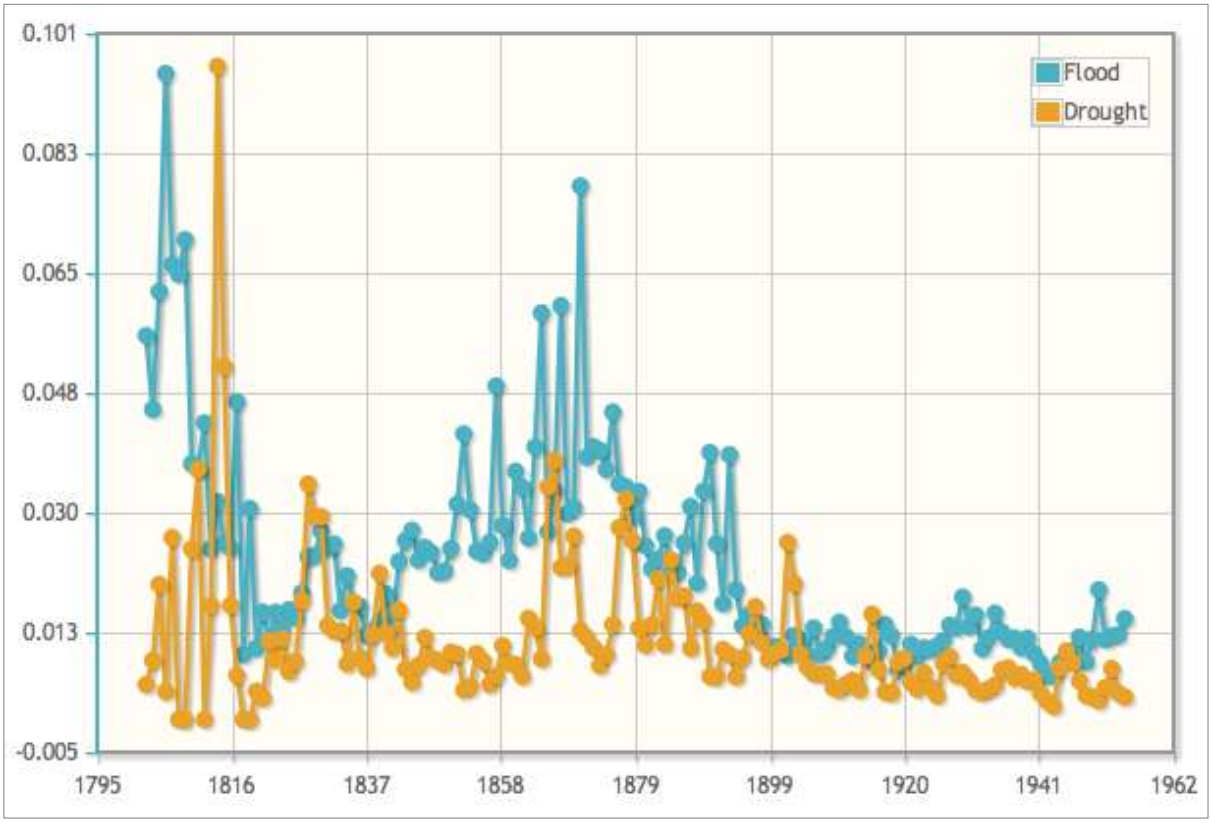

Figure 1: Proportion of newspaper articles in Trove containing the word 'flood' versus the proportion containing the word 'drought'.
http://wraggelabs.com/shed/trove/graphs/flood_drought.html

Like most visualisations, these graphs are intended to raise questions, rather than provide answers. They help you to focus your existing investigations or alert you to a possible new avenue of attack. To encourage further exploration, the graphs are themselves interactive. Clicking on any point will retrieve and display a sample of the matching articles for that year. From the overview you can quickly drill down to individual articles, pursuing the context and quirks of usage. Each graph provides a new, custom-made interface to Trove itself.

There is nothing particularly innovative in graphing search results over time, though the ability to compare the frequency of words and phrases does offer a mode of exploration that is both intriguing and fun. The Google Ngram Viewer (2011) provides a means of visualising queries across the contents of millions of digitised books. Bookworm (2011) goes further, with thematic queries and the ability to drill down to

individual titles. The point is that relatively simple means of analysis and visualisation can allow you to engage with large cultural datasets in new ways.

What other interfaces might be possible? The article previews are retrieved via an 'unofficial API' that I built on top of my screen scraper and host on the Google App Engine (Sherratt 2011e). It is intended for experimentation, to help developers and researchers think about how they might use, access and manipulate data from the Trove newspapers database. The API was used by Newserve (2011), an experimental interface to Australian newspaper collections developed for the 2011 LibraryHack competition.

All of my scripts, tools and examples are freely available under open-source licences (Sherratt 2011f). Anyone can download, play, extend, and improve them. My intention is not merely to service the needs of other historians, but to engage with them in an ongoing conversation about what might be possible in this age of digital abundance.

The times of our lives

I have the sources, I have the tools… now what? I want to use them to explore time.

In my PhD thesis, 'Atomic Wonderland' (Sherratt 2003), I examined how the idea of progress is bound to our conceptions of time. Just as time 'marches on', so we imagine progress as a journey ever onwards; there is no turning back. Progress is catalogued as a succession of novelties, of breakthroughs and revolutions, that separate us from the old world left trailing in our wake.

However, our lived experience is something different. Our lives are full of echoes of the past and plans for the future, memories and imaginings, hopes and regrets. I am hoping that the Trove newspaper database will help me to explore this complexity on a broader level: to see the past not as a series of discrete, chronological events, but as a complex network of comparisons, connections and continuities.

The 'Times of our lives' project, supported by the National Library of Australia through the Harold White Fellowship scheme, aims to open up this expanded temporal space to navigation and discovery. I am planning to undertake three concurrent studies to examine:

- possibilities for extracting events from newspaper text
- historical and cultural practices around the naming of temporal events
- the expression of change and continuity over time

The first study will experiment with existing named entity recognition and event extraction software, such as GATE, to see how effectively events can be identified within the content of newspaper articles.

But what is an event? The second study will investigate the way we name and organise temporal features. We are all familiar with phrases such as 'the Roaring Twenties' but, as Jason Scott Smith (1998) argues, the idea that decades can be packaged up and labelled itself has a history. I want to understand better the processes of periodisation, to know when certain labels were applied, when their meanings changed, and when they fell out of fashion. This information can then be fed back into the Named Entity Recognition process.

The final study will try to elaborate on the continuities of our temporal experience by examining our fondness for dividing lines. According to Elizabeth Eisenstein (1966 p.59), we inhabit ‘history book time’, assigning the past to a series of sequential episodes only to have the narrative break off at the ‘most personally significant, densely-packed, fact-crowded final chapter’. As a consequence, each successive generation imagines itself at ‘great divide’, at a major turning point in history. And so the past is littered with discarded crossroads and forgotten revolutions, with new worlds and new ages that never quite burst into life.

The break between past and present, old and new, is central to the idea of progress. By seeking out failed turning points and examining them against the currents of continuity that run through our daily lives I hope to understand better the power of ‘new’ to shape our priorities and expectations.

These two latter studies will be largely exploratory. Using my existing tools I will follow my hunches, sampling and sketching as I go. Each graph will be an opportunity to learn: about my topic, about the newspapers, and about the search interface itself. For example, a graph charting the phrase ‘new age’ included matches for things like ‘new cordage’, even though I had disabled fuzzy matching. This, I learnt, was due to the way hyphenated words are indexed. The system is of course designed to maximise possibilities for discovery, but this can cause problems when we are trying to track exact phrases across time.

More often, I am likely to be reminded of the dangers of making assumptions. The same graph of ‘new age’ shows a distinct peak in 1941. My first thought, that this indicated people were starting to think about what would come after the war, was quickly dispatched. Retrieving a few sample articles I discovered that the government had introduced ‘new age groups’ for military service. Such are the lessons of the eager text miner.

This will clearly be an iterative process, moving from big picture to individual article and back as I explore contexts, ask questions and test ideas. In one early experiment, I have started to prod at the workings of periodisation by investigating when ‘the Great War’ became ‘the First World War’ (Sherratt 2011d). At some point we realised that the Great War was not the final act in a centuries-long drama of European jealousy and jostling, but the first in a series of global conflicts. Can newspapers tell us when?

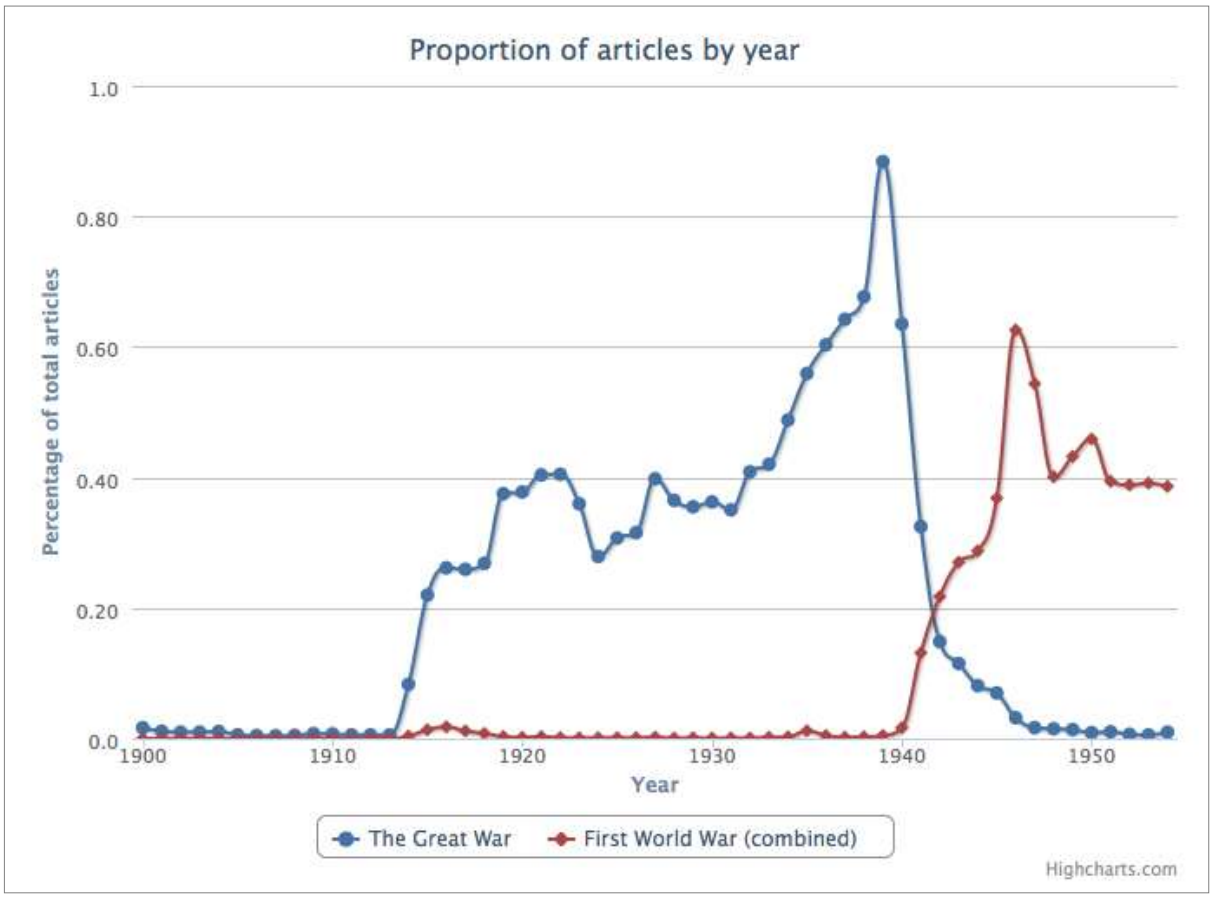

Figure 2: When did the 'Great War' become the 'First World War'? http://wraggelabs.com/shed/time/the_great_war-2011-08-16.html

A graph plotting comparative frequencies shows the two lines crossing late in 1941. With German victories across Europe and North Africa, the opening of the Eastern Front and, finally, the Japanese attack on Pearl Harbor, 1941 seems to make sense. Nonetheless, it is interesting to see this reflected so clearly in such a rough and ready analysis.

What is perhaps more intriguing is a huge spike in 1939. It makes sense that people would be referring back to the Great War as the prospect of a new conflict loomed, but it does make you wonder about the context of these discussions and how they might have developed as war edged closer. As I noted, the value of these sorts of analyses lies in the questions that they raise.

Texts and contexts

What is it that we are actually searching? As we dive into the deep end of massive historical resources like Trove, still we have to maintain a critical distance. These collections are constructed and, like the tools we use to access them, contain arguments about significance, relevance and meaning.

In part, this is familiar territory for the historian, who is used to interrogating the biases of source material. We do not expect newspapers to offer neutral renderings of the past. We understand their political leanings and different styles of reportage. The hit and miss process of discovery in the analogue world also alerts us to the contingencies of creation and preservation. We are too familiar with the experience of 'not finding' to take any source for granted.

But once collected, digitised and delivered on demand, such resources can acquire an air of completeness and finality. Just as we take for granted Google's power to find whatever we want, so our horizons can become constrained to the world within our browser.

Trove will never contain a complete record of our newspapers. Perhaps more significantly, Trove will never be finished. The newspaper database is constantly changing, not only because new articles are being added, but also because volunteer correctors are busily improving the quality of the OCR-rendered text. A search repeated after a few months' interval might return significantly different results (see Figure 3).

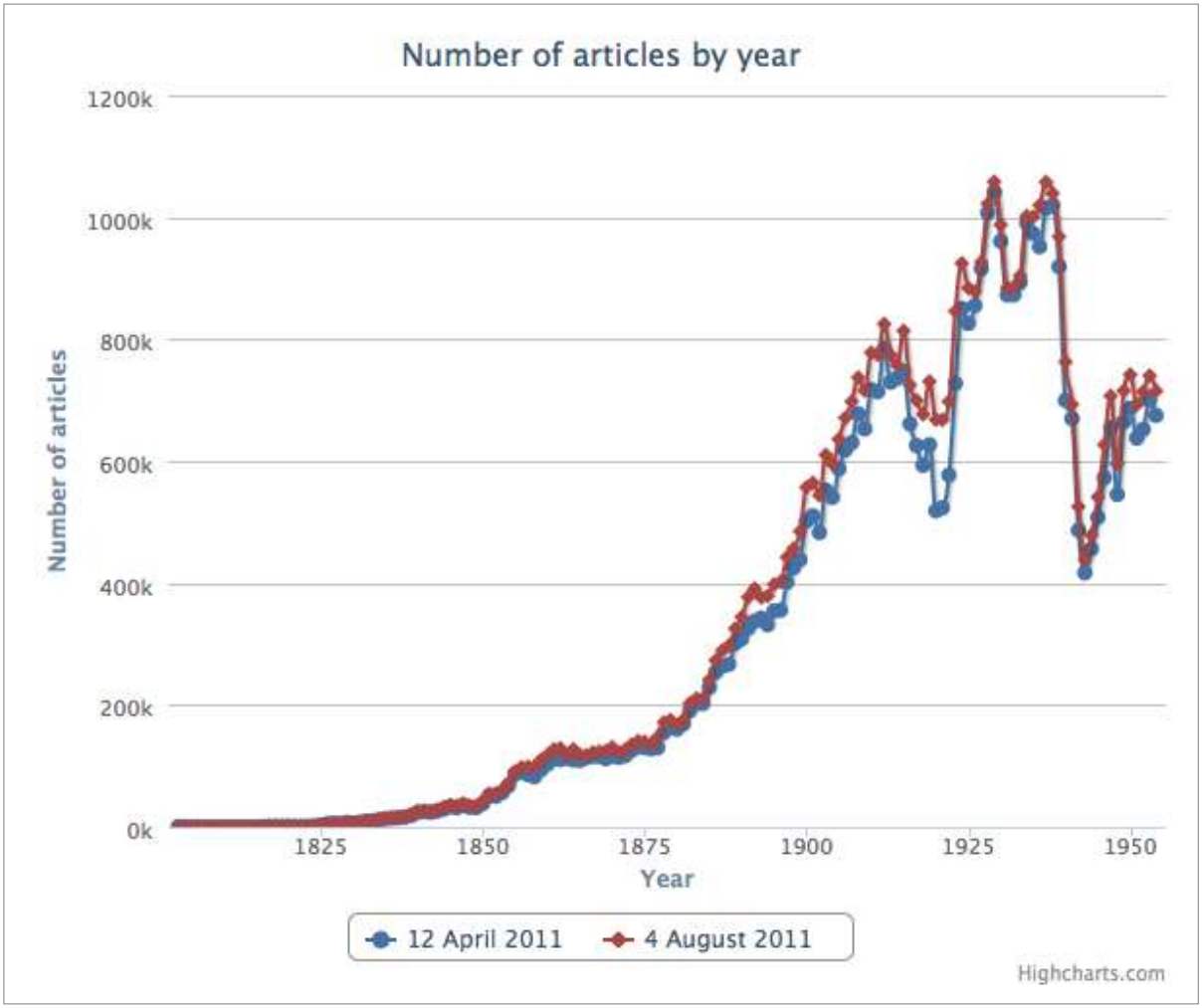

Figure 3 Total number of newspaper articles in Trove. http://wraggelabs.com/shed/trove/graphs/summary_totals.html

This is a good thing — we want more, we want better! However, an understanding of how these processes work and what they mean needs to become part of our interpretative toolkit.

To provide some context for my analyses, I have created a few graphs that represent the complete holdings of the newspaper database. Their value is limited, but they do at least give a quick appreciation of where the collection's strengths are, and a feeling for its evolution. I have also incorporated date stamps into my various tools,

so that researchers have a record of when a particular harvest was made or graph produced. You can easily re-harvest a query to see how results change over time.

Working with large historical resources demands a reflexive turn of mind. As we explore the limits and possibilities we will imagine new interfaces, new tools, and new modes of access. At the same time, we will be forced to re-examine our scholarly practices. The question of how we cite or share an analysis of a resource that is constantly changing, for example, demands an answer that combines both technical innovation and cultural change.

Trevor Owens and Fred Gibbs (2011) suggest we need new modes of writing within the digital humanities that explicitly describe and critique our methods. As we try and understand how the coming age of digital abundance will change the meaning of research, we need to share more than our results: we need to show our working out.

Equally we need to find space for new modes of engagement, based not just on the constant search for sources, but on experiment and play. Stéfan Sinclair (2003 p.181) argues that 'play is an integral part of a humanist's interpretive activities'. We need to both recognise play as legitimate scholarly activity and create tools that encourage it.

This paper has explored some of the ways in which new digital technologies can allow us to explore and engage with large cultural resources such as the Trove newspapers database. These tools and technologies enable us to move beyond standard discovery interfaces and grapple with growing abundance of online sources. We can extract and manipulate structured data, we can work at a variety of scales, and we build our own access points and visualisations. The point is not to build a better search engine, but to support new understandings and analyses.

As Father Roberto Busa reflected upon his lifelong labours he recognised that the value of computers lay not in the fact that they made things faster or easier:

> … the use of computers in the humanities has as its principal aim the enhancement of the quality, depth and extension of research and not merely the lessening of human effort and time. (Busa 1980 p.89)

Digital resources enable us to do new things, to think in new ways, to ask new questions and suggest new answers. I am interested in mining Trove, not as a technical exercise, or merely as an enhanced mode of discovery. I am interested in the possibilities it offers for the creation of new knowledge, for learning about ourselves.

References

Allen, R.B. & Sieczkiewicz, R., 2010. How historians use historical newspapers. *Proceedings of the American Society for Information Science and Technology*, 47(1), pp.1-4.

Anon, 1828. Newspapers *The Australian*, p.4. Available at: http://nla.gov.au/nla.news-article36867029 [Accessed October 22, 2011].

Anon, 1905. The educational value of “news.” *The State*.

With Criminal Intent, 2011. Available at: http://criminalintent.org/ [Accessed November 5, 2011].

Bagnall, K., 2011a. Going against the grain. *The tiger’s mouth*. Available at: http://chineseaustralia.org/archives/1140 [Accessed November 5, 2011].

Bagnall, K., 2011b. Rewriting the History of Chinese Families in Nineteenth-Century Australia. *Australian Historical Studies*, 42, pp.62-77.

Bookworm, 2011. Available at: http://bookworm.culturomics.org/ [Accessed January 26, 2011].

Busa, R., 1980. The annals of humanities computing: The index Thomisticus. *Computers and the Humanities*, 14, pp.83-90.

Dalton, M.S. & Charnigo, L., 2004. Historians and their information sources. *College & Research Libraries*, 65(5), p.400.

Cohen, D.J., 2006. From Babel to Knowledge. *D-Lib Magazine*, 12(3). Available at: http://www.dlib.org/dlib/march06/cohen/03cohen.html [Accessed July 30, 2011].

Cohen, D.J. et al., 2008. Interchange: The Promise of Digital History. *Journal of American History*, 95(2). Available at: http://www.journalofamericanhistory.org/issues/952/interchange/ [Accessed November 2, 2011].

Cohen, D. et al., 2011. *Data Mining with Criminal Intent*, Available at: http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf.

Eisenstein, E.L., 1966. Clio and Chronos: An essay on the making and breaking of history-book time. *History and Theory*, 5 (Beiheft 6), pp.36-64.

Gibbs, F.W. & Owens, T., 2011. Hermeneutics of Data and Historical Writing. In J. Dougherty & K. Nawrotzki, eds. *Writing History in the Digital Age*. Available at: http://writinghistory.trincoll.edu/data/hermeneutics-of-data-and-historical-writing-gibbs-owens/ [Accessed November 5, 2011].

Google Ngram Viewer, 2011. Available at: http://books.google.com/ngrams [Accessed January 26, 2011]

Hitchcock, T., 2008. Digital searching and the re-formulation of historical knowledge. In Mark Greengrass & Lorna Hughes, eds. *The Virtual Representation of the Past*. Farnham, UK: Ashgate, pp. 81-90.

Hockey, S., 2004. The History of Humanities Computing. In S. Schreibman, R. Siemens, & J. Unsworth, eds. *A Companion to Digital Humanities*. Oxford: Blackwell.

Available at: http://www.digitalhumanities.org/companion/ [Accessed November 5, 2011].

Jones, A., 2009. *The many uses of newspapers*, Tufts University. Available at: http://dlxs.richmond.edu/d/ddr/docs/papers/usesofnewspapers.pdf [Accessed October 22, 2011].

LibraryHack, 2011. *Newserve*. LibraryHack 2011. Available at: http://libraryhack.org/2011/05/31/newserve/ [Accessed November 5, 2011].

Mallet, 2011. Available at: http://mallet.cs.umass.edu/ [Accessed January 26, 2011].

ScraperWiki, 2011. Available at: http://scraperwiki.com [Accessed November 5, 2011].

Sherratt, T., 2003. Atomic Wonderland: Science and Progress in Twentieth Century Australia. PhD. Australian National University. Available at: http://discontents.com.au/shoebox/history-of-australian-science/atomic-wonderland.

Sherratt, T., 2011a. Mining the treasures of Trove (part 1). *discontents*. Available at: http://discontents.com.au/shed/mining-the-treasures-of-trove-part-1 [Accessed November 5, 2011].

Sherratt, T., 2011b. Mining the treasures of Trove (part 2). *discontents*. Available at: http://discontents.com.au/shed/experiments/mining-the-treasures-of-trove-part-2 [Accessed November 5, 2011].

Sherratt, T., 2011c. Some exhibition magic with Zotero and Omeka. *discontents*. Available at: http://discontents.com.au/shoebox/weather-research-topics/some-exhibition-magic-with-zotero-and-omeka [Accessed November 5, 2011].

Sherratt, T., 2011d. When did the “Great War” become the “First World War”? *discontents*. Available at: http://discontents.com.au/shed/experiments/when-did-the-great-war-become-the-first-world-war [Accessed November 5, 2011].

Sherratt, T., 2011e. *Australian Newspapers API.* Available at http://wraggelabs.appspot.com/api/newspapers/ [Accessed November 5, 2011].

Sherratt, T., 2011f. *WraggeLabs Emporium.* Available at http://wraggelabs.com/emporium/ [Accessed November 5, 2011].

Sherratt, T., 2011g. It’s all about the stuff: collections, interfaces, power and people. *discontents*. Available at: http://discontents.com.au/words/conference-papers/it%E2%80%99s-all-about-the-stuff-collections-interfaces-power-and-people [Accessed January 26, 2011]

Sinclair, S., 2003. Computer-Assisted Reading: Reconceiving Text Analysis. *Literary and Linguistic Computing*, 18(2), pp.175 -184.

Smith, J.S., 1998. The Strange History of the Decade: Modernity, Nostalgia, and the Perils of Periodization. *Journal of Social History*, 32(2), pp.263-285.

Tibbo, H., 2003. Primarily History in America: How U.S. Historians Search for Primary Materials at the Dawn of the Digital Age. *American Archivist*, 66(1), pp.9-50. Available at: http://archivists.metapress.com/content/b120370l1g718n74 [Accessed July 14, 2009].

Turkel, W.J., 2005. Teaching Young Historians to Search, Spider and Scrape. *Digital History Hacks (2005-08)*. Available at:

http://digitalhistoryhacks.blogspot.com/2005/12/teaching-young-historians-to-search.html [Accessed September 7, 2011].

Turkel, W.J., 2006. Methodology for the Infinite Archive. *Digital History Hacks (2005-08).* Available at: http://digitalhistoryhacks.blogspot.com/2006/04/methodology-for-infinite-archive.html [Accessed September 7, 2011].

Pettinato, T., 2010. Newspapers: “the rough draft of history.” *Readex Blog*. Available at: http://blog.readex.com/newspapers-the-rough-draft-of-history [Accessed October 27, 2011].

Nelson, W., 1909. The American newspapers of the eighteenth century as sources of history. *Annual report of the American Historical Association for the year 1908,* 1, pp.209-222.

WordSeer, 2011. Available at: http://www.eecs.berkeley.edu/~aditi/projects/wordseer.html [Accessed January 26, 2011].

Zotero, 2011. Available at: http://zotero.org [Accessed January 26, 2011].

Zweig, R.W., 1998. Lessons from the Palestine Post Project. *Literary and Linguistic Computing*, 13(2), pp.89 -94.