Telling stories with data

Author : Tim Sherratt

Tagged as : bots, digital humanities, invisibleaustralians, Linked Open Data, QueryPicTrove

Date : August 25, 2016

Keynote presented at Working History, the Professional Historians' Association Conference, 19 August 2016, Melbourne.

Friends, I come bearing good news. For I have seen the future of history.

Indeed I saw it here in Melbourne in February. There it was in the exhibition hall of the VALA2016 conference…

VALA 2016
The Future of History
83
datacomIT.
THE FUTURE OF History
2 x Platinum Award Winning microfilm scanner with a 14 Megapixel USB 3.0 colour camera
DL mini
Books

THE FUTURE OF HISTORY

…is a very fancy book scanner.

It’s easy to get caught up in the hyperbole – to imagine that history has been swept up in a technological revolution.

I must admit, I’ve often argued that access to more than 200 million digitised newspaper articles through Trove has profoundly changed the practice of history. I still believe that. It’s not just about convenience – the fact that you can now do your research at home in your pyjamas. It’s also about making the fragile slivers of ordinary human experience accessible in a way that just was not possible before. It’s given us the ability to tell different types of stories.

But is it a revolution? As historians I think we have an obligation to be sceptical of the ‘R’ word – we all know that for every change there is a continuity. And that’s what interests me in the digital space, the interplay between historical practicisibe and technology; between new possibilities and old critiques.

Fear not – you will not be rendered obsolete by army of sentient book scanners. But perhaps there are different ways we can work, different questions we can ask.

Discovery against the grain

So, Trove.

I’m assuming in audience such as this I don’t need to explain what Trove is. Of course Trove is much, much more than just digitised newspapers, but let’s just focus on those newspapers for a minute. What we’re talking about is:

- More than 200 million newspaper articles from 1803 onwards
- More than 1000 different newspaper titles – not just the metropolitan dailies; rural and regional papers; political, religious, and community papers; papers in a range of different languages.
- Terabytes of text – all searchable

I think we tend to take this last point for granted – not only can you explore 200 million newspaper articles online, but you can search inside them. I think there is real revolutionary force in the combination of OCR (Optical Character Recognition) and keyword search. It shifts power away from the headline writers, the cataloguers, the archivists and the indexers, and opens an infinite number of pathways for discovery. I’m sure all of you have a story about some tiny fragment, a clue that was critical to your research, buried deep in a seemingly uninteresting newspaper article. This was always possible if you had enough time and physical access to newspapers, but technology has normalised this mode of exploration. We are no longer fossickers, scouring past workings in the hope of finding a gem. Whether we know it or not, we now all dig deep along unmapped seams of history and meaning.

But this is only one example of how discovery is moving beyond metadata – the information we record about resources – to mine the very content of those resources, to create new access points. Newspapers are easy, what about handwritten resources like letters and diaries? Technology has enabled new forms of collaboration between institutions and researchers, allowing them to create and share transcriptions of otherwise unsearchable documents.

As of last week, for example, volunteers had transcribed more than 16,000 manuscript pages from the papers of English philosopher and reformer Jeremy Bentham. What’s particularly interesting about the Transcribe Bentham project, is that the transcriptions are themselves being used to train new tools for automated handwriting recognition. Humans are teaching machines, who will in turn help humans to open resources to new forms of discovery. There’s still a way to go, but the possibilities are pretty exciting. What will happen when instead of consulting an index to a large collection of correspondence, we can search the complete content?

And it’s not just text.

Cultural institutions have been experimenting with colour as a way of navigating large image collections. The colour values of the images themselves are extracted, aggregated and normalised, allowing users to discover connections otherwise unknown or undocumented. Companies like Google are using artificial intelligence to generate descriptions of images. This technology might enable us to search millions of uncatalogued photographs held by cultural heritage institutions. Like switching on the lights in a darkened room, the application of computing power will help reveal stories lurking in the shadows of anonymity.

Assuming of course that they have been digitised.

The seeming omnipotence of Google has encouraged us to equate discoverability with existence – if we can’t find it, it doesn’t exist. But only a small proportion of our cultural heritage collections have actually been digitised. Even as we take advantage of these new technologies we have to think about what’s missing. Whose history has been digitised, described and processed, and why?

This graph shows the number of digitised newspaper articles in Trove by year. As experienced, professional historians you will of course know why there is a peak around 1914. More news…? More newspapers…? The answer is more money for digitisation. With the centenary of World War I approaching, it was decided to focus resources on the digitisation of newspapers from that era. This is completely understandable, but also largely invisible to users.

At the recent digital humanities conference in Hobart, Tim Hitchcock talked about the hidden histories that shape the online resources on which we now depend. We are working with collections of text, he argued, that are ‘inherently, and institutionally, Western centric, elitist and racist’. With dollars for digitisation becoming increasingly scarce we can expect more ‘project based’ funding responding to particular events or anniversaries. Will this simply reinforce the limits of our digital horizons?

We need to understand that we start our digital explorations from a point of privilege and exclusion. Decisions have been already been made about what and who matters. But within these limits we can continue to work against the grain, to flip perspectives, to see things differently.

About five years ago I worked out how to do bulk downloads of digitised files from the National Archive of Australia’s online database RecordSearch. My partner Kate Bagnall, a historian of Chinese-Australia, and I were particularly interested in the many thousands of records they hold documenting the workings of the White Australia Policy. So I downloaded about 12,000 pages, many including portrait photographs, and ran them through a simple program to find faces. The result was a scrolling wall of about 7,000 portraits that we called ‘The Real Face of White Australia’ . The remnants of a racist bureaucratic system were turned inside out – instead of files and metadata you could see the people inside.

It’s these sorts of possibilities for seeing against the grain that get me most excited about digital technologies. We no longer have to accept what we’re given by cultural institutions, we can build and share new perspectives, even new interfaces.

One of my current projects is exploring the records held by the National Archives of Australia that we’re *not* allowed to see – those with an access status of ‘closed’. Anyone who’s used the National Archives knows that the access examination process can sometimes be a bit frustrating. But it *is* just a process – there’s nothing magical or mysterious about crossing the threshold from closed to 'open'.

In this case there's obviously no files for me to download, but there is data documenting when and why a file was closed. By harvesting that data from RecordSearch and feeding it through a new interface we can start to build up a picture of how that process works – we can investigate access itself as a historical phenomenon.

Working at scale

More than a decade ago, the pioneer digital historian Roy Rosenzweig talked about the challenges of digital abundance – how would historical methods change to deal with the vast quantities of digital content becoming available?

What does it mean when you search for something in Trove and find you have 10,000 matching results, or maybe 100,000, a million? We're used to working in a linear fashion, interrogating our sources one at a time. So how can we extract understanding from a set of resources that we can never hope to examine individually?

It was this question that inspired me to create QueryPic, a simple tool for visualising searches in Trove's digitised newspapers. You enter your keywords in the usual fashion, but instead of a list of search results you're presented with a chart that shows you the number of articles each year that match your query. Instead of just the top twenty results, you see everything at once. Using QueryPic you can combine multiple queries to track changes in language or technology, or observe the impact of particular events. (In this case we're comparing the use of the name 'Santa Claus' vs 'Father Christmas'). It's easy to create, save and share charts – have a go!

There are many similar tools around these days. Google's Ngram viewer enables you to construct complex queries across the contents of millions of books. Bookworm lets you explore trends in a range of sources including US newspapers and the scripts of the Simpsons.

Most of these tools work by treating texts as data – by breaking texts down into their component parts, and then analysing the occurrence of particular words, phrases, or other patterns. Digital historians are lucky. Literary scholars have been working for decades on the computational analysis of text, and the tools they've created can be readily applied to historical sources.

Voyant Tools, for example, is a web-based text analysis platform. Feed it text files, web pages, XML, even PDFs, and it will slice and dice the language of your sources, opening the results for further exploration through a series of interactive tools. It's powerful enough to handle many megabytes of data. In this it's analysing the talk that I'm giving right now.

If you want to start somewhere simpler, have a play with DataBasic.io, where you can learn the fundamentals of text analysis by diving deep into the lyrics of Beyonce.

QueryPic generates large scale pictures using Trove's digitised newspapers, but what if you want something more fine-grained? I've also created a Trove Harvester that will save the details of all newspaper articles matching a particular query to your own computer. Perhaps you want to look for patterns in the language of all articles that include the phrase 'White Australia'. Just grab the contents of the articles using my harvester and upload them to Voyant. Bam!

This form of analysis is often termed 'distant reading'. Instead of examining individual documents we use computational methods to look for patterns across a large collection of documents. By zooming out, we can explore the historical record at different scales, finding new connections and meanings.

Have you come across the Old Bailey Online? It's an astonishing resource, providing fullysearchable text of nearly 200,000 criminal trials from 1674 to 1913. But of course it's not just text. The data is structured so you can search by name, crime, verdict and punishment. Zooming out of individual trials, historians can examine changes in the way the legal system itself operated over time. Similarly, the Prosecution Project is compiling data about Australian criminal trials from a variety of sources, including Trove. Who knows what we might learn about the nature of our criminal justice system.

The data gathered by projects such as these allow distant reading not just of language, or institutions, but of populations. Who were these people whose lives intersected with the administration of the law and the operations of the state?

The Digital Panopticon project is now linking up records from a number of these different databases to track people from the Old Bailey, through transportation to Australia, and beyond.

One of the things I love about this sort of work is that the data is so richly and profoundly human. In this age of so-called 'big data' there's a tendency to imagine that a focus on computational methods will somehow firm up the scientific credentials of the humanities. We have data too! Indeed, Google's Ngram viewer was announced to the world as as the flag bearer of a new field called 'culturomics' that would bring statistical rigour to the historical study of language and culture.

Yes, it's bullshit. Big data is made up of many small acts of living. And life, as we know, is messy and complicated, and resists our attempts at categorisation. That's what makes history so much fun. As historians we have the opportunity, and indeed the obligation, to tease out the connections between the micro and the macro; between the unkempt trajectories of individual lives and the beautiful curves of our data visualisations.

Click any point on a QueryPic chart and you'll see the first twenty matching results from Trove. QueryPic's visualisations are not arguments, they're starting points – ways of exploring ideas, or surveying new territory. Understanding comes from shifting scales – moving between individual articles and long term trends.

One of the things that excites me most about digital tools and techniques are the opportunities they create for navigating these changes in scale. We can create new resources and interfaces that bring together statistics and stories; that enrich our data with the power of narrative, and vice versa.

Doing it in public

A couple of years ago I started to collate information about websites that included links back to digitised newspaper articles on Trove. I wanted to understand more about the contexts in which the newspapers were being used and cited. The diversity of subjects and sites was astonishing, and sometimes disturbing. But what was particularly interesting was the amount of historical work some people were doing.

KnowThatProperty.com sounds a bit like a commercial real estate site, but in fact it provides potted histories of houses around Sydney, largely drawn from Trove. For example, the entry for number 2 Carrington Street in Strathfield, provides details of the house's construction and ownership from 1888 to 1927, with over 40 links to newspaper articles in Trove. The creator of the site is a web developer with an interest in architectural history. But is he a historian?

Who cares?

Ready access to historical sources through services such as Trove allow people to pursue their passions, around and beyond the demands of everyday life. We are no longer subject to the tyranny of the microfilm reader. The *work* of history – the chasing down of connections, the exploration of context, the compilation of references – is no longer confined to designated places of research. Nor is it expressed solely in traditional forms of historical production.

Digital tools help us find things, but they also help us share them. We write blog posts, we collect on Pinterest, we repost on Tumblr, we ‘like’ on Facebook. You might think that this is all a bit trivial, and sometimes it is, but the dynamics of sharing help us to look at history differently. The ‘public’ are no longer external to the process of history making – an imagined audience, or prospective consumers. We’re just all in there together.

Who’s made a list on Trove? Trove lists are just collections of interesting items. You create lists on particular topics and use them to keep track of relevant resources. You might create and share and share a list of newspaper articles relating to your family’s history for instance. But, once shared, a list becomes something more than a convenient bucket of content. Lists provide thematic entry points that aid discovery by creating implicit links between items. They’re also building blocks for new forms of access.

Last year I created a web application that takes the contents of Trove lists and turns them into online exhibitions. Using freely available web services, you can create your own exhibition in minutes (yes, really).

I’m sorry, but I really don’t care about questions of authority or professional identity. The guy who created KnowYourProperty.com isn't waiting for the historians of Australian to pin a membership badge on him. Nor, I suspect, is the bloke who has created more than 200 Trove lists about lawnmowers worried about whether what he's doing is really history. I happen to think it is, but more importantly it's about passion. It's about *just doing something*. In a world that champions consumption over creation, conflict over collaboration, we should celebrate any effort to make an authentic connection to the past.

There's something quite liberating about the fluidity of the digital environment. Instead of trudging the well-worn path from research to product we can explore the possibilities of reuse; we can experiment with form and meaning; we can play; and we can feel.

The Vintage Face Depot is a Twitter bot. Tweet a picture of yourself to it and you will receive back a modified you – your face will be blended with one drawn at random from a collection of faces extracted from Trove's digitised newspapers. It sounds creepy and I suppose it is, but the effect can be quite interesting, sometimes unsettling. What happens when you see your own eyes peering out of a face from the past? The image is accompanied by a link to the original article on Trove, so you can find out more about who you've been blended with. Like a lots of my work, the Vintage Face Depot was made quickly as an experiment. I still don't know what to think of it. And that's good.

Caleb McDaniel's Twitter bot, @every3minutes is a more deliberate intervention in our experience of the past. Historians of the American slave trade have estimated that a person was sold every three minutes between 1820 and 1860. So Caleb's bot tweets, every three minutes – 'someone just purchased a black person's grandchild', 'a white slaver just sold a person's friend'. It's unrelenting and confronting.

We are finding new forms of historical expression that don't merely visit the digital realm, they live there.

Show your working out

Earlier this year I sort of accidently created and shared my own version of Commonwealth Hansard . It's a site that lets you browse the proceedings of the House of Representatives and the Senate from 1901 to 1980. I was originally intending just to harvest the data that sits underneath Hansard on the Australian Parliament House website. I thought all that nicely-structured text would provide an interesting dataset to poke around in using the sorts of text analysis programs I've already described. But after downloading about 4gb worth of data, I starting thinking about other things I could do with it.

If you've used Hansard on the Parliament site, you'll know that it's hard to read anything in context – you're constantly navigating your way up and down a confusing hierarchy of debates and speeches. So I decided to make a version of Hansard that was focused on reading – one sitting day per page. That's it.

But that simplicity has allowed me to do other things. Every year, and every sitting day, has a button that automatically opens the proceedings for that period in Voyant. That simple button turns a page of text into a laboratory for the historical analysis of political speech. I've also integrated Hypothes.is which allows anyone to annotate the text – adding notes, links, highlights, even images. Annotations created with Hypothes.is can be shared globally, turning each page of text into site for collaborative research.

Both Voyant and Hypothes.is can be added to any web page with just a couple of lines of HTML code. It makes you wonder why we are still recreating traditional forms of publication online, when we could be doing so much more, so easily.

You would have noticed that this set of slides itself embeds a number of graphs, visualisations, and live web pages that you can play with inside the presentation. The creators of Voyant, Stefan Sinclair and Geoffrey Rockwell, believe it is important to create analytical tools, or hermeneutica, that can be embedded within works of scholarly interpretation. Voyant's widgets encourage readers to play with the data and not merely consume the argument. They give power to readers to build their own interpretations, to make their own discoveries.

In a similar way, the historian Tom Griffiths has described footnotes as 'generous signposts to anyone who wants to retrace the path and test the insights'. Footnotes too give power to the reader, but in a non-digital environment the ability to exercise that power is deferred, perhaps indefinitely. How can make sure those 'generous signposts' actually point somewhere?

Perhaps you've heard of a thing called Linked Open Data – it's really just a way of publishing nicely structured data on the web so that it can be easily connected up across sites and collections. Historians create Linked Open Data all the time, they just don't know it. Think about all those spreadsheets or index cards you have listing people, linking them to other people, places, events, and documents. In LOD terms these are all nodes and edges – entities and relationships.

Historical research frequently involves creating these sorts of complex data models. They represent a huge investment of skill, knowledge and experience. But what happens when we come to 'write up' our research? The data is squeezed out of the narrative, flattened down to comply with the conventions of linear storytelling. The connections are severed.

Kate and I are playing around with ways of combining historical narrative and Linked Open Data to make sure that the story remains in conversation with the data – to give readers the freedom to jump off at any time into the underlying network of people, places, events, and resources. And if those people, places, events, and resources are themselves linked to the holdings of libraries, archives, and museums, then every piece of writing becomes a gateway – a starting point for further exploration of our cultural collections. Generous signposts indeed.

Our latest LODBook experiment is available online. It's buggy and incomplete, but gives a sense of what we want to do. I'd originally intended to have a nice, finished version to show you today, but I decided recently to throw out a lot of code and start again. Why? I always wanted something that was simple and sustainable – a set of practices and reusable components, rather than yet another publishing platform. Having been recently inspired by work going on around the idea of minimal computing, I decided I needed to focus more on the basics. So watch this space.

Digital tools and technologies give us the opportunity to experiment, and sometimes to fail. As I was harvesting text from Hansard I noticed a few oddities. After further investigation I realised that more than 90 Senate sitting days were missing from Parliament House's online database, including about half of 1917. It's unlikely anyone would have noticed the problem using the web interface unless they were looking for a specific date. Historians researching the World War I period just don't know what they've missed. Fortunately the people at Parliament House are now investigating to see what they can do.

These things happen. Systems are never perfect. But when they do happen it's important to talk about them. This is a great example of why we need to remain critical of search as a means of access – it can't find what's not there.

The story of the Senate black hole is documented in my open research notebook – it's where I post notes and experiments relating to my current research projects. It's an idea I've stolen from a number of colleagues working in digital history, including Caleb McDaniel, and I think it's a good example of how the digital environment can encourage us to re-examine our practices.

The non-digital world privileges products as markers of achievement – things we can count, things we can launch, things we can sell. A conference on ‘working history’ seems like an ideal place to challenge that, to think about how we can use digital tools and techniques to expose more of the labour, the craft, the practice of history. To focus on the doing, not the done.

In 2008, the American historian William G Thomas suggested that ‘digital history should embrace the impermanence of the medium, use it to convey the changing nature of the past and of how we understand it’. The digital future is full of clamour and distraction, an overwhelming array of possibilities. Much like the past.

Our experiments, our incomplete thoughts, our works-in-progress, our failures, reflect the confusion and uncertainty we can never escape. By exposing them online, we are simply admitting what we’ve always known. History is constantly in the process of being made.

Share this:

- Click to email this to a friend (Opens in new window)
- Click to print (Opens in new window)
- Click to share on Twitter (Opens in new window)
- Click to share on Facebook (Opens in new window)
- Click to share on Google+ (Opens in new window)
-

This work is licensed under a Creative Commons Attribution 4.0 International License.