Asking better questions

History, Trove and the risks that count

Tim Sherratt

A few years ago historian Kate Bagnall and I created 'The real face of White Australia'. In 1901 the *Immigration Restriction Act* gave legislative force to a system of racial exclusion and control that came to be known as the White Australia Policy. The bureaucratic remnants of this system survive today in the National Archives of Australia. But how can we find them? How can we see them? Our online experiment brings some of these lives, previously deemed out of place in a 'white' Australia, to the surface. Instead of documents, files or search results, all you see are faces – a continuous, scrolling wall displaying thousands of faces. It's compelling, challenging and discomfiting. Some viewers were brought to tears.

The faces come from portrait photographs that were attached to official certificates. If non-white residents wanted to travel overseas, they needed special identity documents. Without them they could be refused entry on their return. They would not be allowed to come home. On the front of each certificate are photographs

and basic biographical details, on the back is a palm print. Many thousands of these documents are preserved in the archives.

To extract the faces I reverse-engineered the National Archives' online database to automatically harvest images of the documents. From just one series or group of records, I downloaded more than 12,000 images. Then I tinkered with some facial detection code until I was able to find and crop out the portraits. I ended up with 7000 faces – just a sample of the archives' holdings, but enough.

The whole thing was a quick experiment, mostly completed over the space of a weekend, but its influence has been widely felt. The project has been cited in discussions around race, archives, visualisation and serendipitous discovery. It has been assigned for student reflection in university courses around the world and is regularly held up as an example of what the digital humanities has to offer. But whenever I give a talk about it in Australia, one question seems inevitably to arise – what is the copyright status of the images?

In a keynote address to New Zealand's National Digital Forum in 2011, library technologist Michael Lascarides challenged those of us who work with digital cultural collections to 'ask better questions'. When confronted with the remains of our racist history, when looking into the eyes of people whose lives were monitored and controlled by the government because of the colour of their skin, why should we feel compelled to consider the technicalities of copyright?

Surely there are better questions to ask?

©

I describe myself as a cultural data-hacker, but my business hours are currently spent on the other side of the fence, as the manager of Trove at the National Library of Australia. Trove is a discovery service that makes Australian resources easier to find and use. It's a collection of collections, bringing together the holdings of many libraries, archives, museums, universities, government agencies and more. The most heavily used part of Trove is an ever-growing collection of digitised newspapers – the full text of more than 130 million articles documenting Australian history from 1803 onwards.

It's hard to measure the cultural impact of Trove's digitised newspapers. Technologies like optical character recognition (OCR) and keyword searching are now commonplace, but apply them to 150 years of Australian history and something transformative happens. Easy access to historical newspapers is changing our relationship with the past.

It's not just about convenience – the ability to do your research at home in your pyjamas – although the significance of opening access to rural and remote communities across a large country like Australia shouldn't be underestimated. It's also about using the granularity of newspapers to expose the local, the particular, the personal and the ephemeral – glimpses of ordinary lives otherwise unrecorded.

Kate is a historian of Chinese Australia, interested in intimate relationships between Chinese men and white women. The people she studies often lived at the fringes

of society and their lives can be difficult to recover. But searching across digitised newspapers she can find shards and fragments, stories of love and loss, full of the vivid, turbulent detail of everyday life. Each shard helps to build a bigger picture, a different view of Australian society and history.

Other researchers have mined the newspapers in pursuit of topics as diverse as invasive species, climate change, poetry and legal history. These uses will multiply as the corpus grows and our tools develop. Like a big telescope or a particle accelerator, digitised newspapers support large-scale fundamental research across a range of disciplines. Old papers have become a site for the creation of new knowledge.

But digitised newspapers are not solely the province of professional researchers. The size and diversity of their content support almost any interest, feed almost any passion. I recently harvested a sample of pages on the web that include links to Trove's newspapers – 3116 webpages containing 13,389 links. We know that family and local historians make heavy use of the newspapers and their efforts are well represented in my sample. But there was more – sport, war, science, politics, architecture, music, art … from popular entertainment to academic treatises, from hateful diatribes to thoughtful reflections, they were all there.

> "Easy access to historical newspapers is changing our relationship with the past.

More surprising than just the range of topics, styles and prejudices is the different ways the newspapers are

used online. In 2013, for example, the local media in Western Australia reported on Cockburn City Council's plan to erect a shark barrier at Coogee Beach. When one councillor expressed doubts, noting the lack of 'serious or fatal shark attacks at Coogee Beach since records commenced in the 1800s', a reader could quickly challenge her comments by citing two Trove newspaper articles that documented local attacks. References to the digitised newspapers are embedded online not just within narratives or compilations, but as commentary and debate. Trove provides a ready source of evidence to test historical claims without lengthy research or the mediation of experts.

Easy accessibility is helping to break down the otherness of the past, allowing it to be mobilised in contemporary discussions. New conversations between past and present are emerging around the digitised newspapers. Trove has launched us upon a massive ongoing experiment in collaborative meaning-making.

And it might never have happened.

©

Years before I was given the job as Trove manager, I was poking and prodding at the interface, trying to extract useful data to use in new applications, and generally making a nuisance of myself. Among the tools I created was QueryPic, a simple way of visualising newspaper searches. It's been through several versions but the principle remains constant – just feed in your search query and QueryPic will create a line chart showing you the

number of newspaper articles per year that match your query.

QueryPic is well used and has even been cited in scholarly articles, but I think its greatest value is as an example of what becomes possible when you make large quantities of cultural content available in digital form. Instead of the normal list of search results, QueryPic shows you trends and patterns. You can observe changes in language, the rise and fall of our cultural obsessions or the impact of major events. When did the 'Great War' become the 'First World War'? Is that jolly Christmas visitor called 'Father Christmas' or 'Santa Claus'? QueryPic helps you see things differently.

One of the most dramatic and unexpected patterns revealed by QueryPic is that Australian history ends in 1954. Who would have guessed? With very few exceptions, Trove's collection of digitised newspapers comes to an abrupt halt in 1954, when the possibilities of the digital age meet the realities of copyright. You want to trace cultural patterns beyond 31 December 1954? Sorry, you're out of luck.

Why 1954? We're currently about halfway through the great AUSFTA culture drought. On 1 January 2005, the Australia–US Free Trade Agreement (AUSFTA) extended the standard period of copyright protection from fifty to seventy years and changed the way photographs are treated. We might have to wait until 2025 before Trove's newspapers can start edging forward, year by year, beyond 1954.

But it's even more complicated than that, as there's no certainty that newspaper articles published before 1955

are out of copyright. To be sure, the National Library would have to investigate any named authors to confirm they all died before 1955. That's simply impossible in a mass digitisation project. Instead the library weighed the copyright risks against the cultural benefit and decided to proceed. If the library had been more cautious, if the risks or uncertainties had seemed too great, Trove would have no digitised newspapers. And we would all be poorer.

These types of judgements are made all the time by cultural organisations wanting to open online access to their collections. Libraries, archives and museums are full of so-called 'orphan' works, whose creators cannot be identified or located. There's no risk-free way of making this content available online.

But these risks are not only assessed and managed, they're passed on to users, who must themselves try and navigate the thicket of copyright law. As use of online collections moves beyond traditional forms of citation into new types of digital aggregation, analysis and annotation, the doubts and complexities accumulate. The price of innovation is increased risk. The only safe course is to do nothing.

©

Why do we put cultural heritage collections online? Is it for the sake of efficiency, preservation, marketing or perhaps an informed citizenry? Usually we fall back on fuzzy notions of 'engagement' or 'access'. More access is good, particularly if we can measure it easily through web stats.

Recently we surveyed Trove users to gain a broad picture of satisfaction and use. One finding in particular keeps me coming into work each day – 90 per cent of our general users agreed with the statement 'Trove has made me interested in learning and discovering more'. Access can't simply be measured in collection images or web-page hits. What we're creating is an enlarged space for reflection, research, learning, creativity and critique. We're enabling people to do more, with more.

Trove is not alone. Around the world, projects such as Europeana, the Digital Public Library of America (DPLA) and Digital New Zealand all work to open our collected cultural heritage to new forms of use. Europeana, in particular, has drawn on its research into the value and impact of online collections to proclaim a wonderfully ambitious agenda. Their aim is to 'transform lives' – to unlock Europe's cultural heritage, enabling it to act as a 'catalyst for social and economic change'.

Resisting Europeana's efforts to 'transform the world with culture' are an array of different copyright regimes across Europe. Advocacy on behalf of the very idea of 'openness' is crucial to the success of their mission. But it's never simply a matter of law.

Our cultural collections contain many resources that are already free of copyright restrictions, but it's not always easy to find them. A lack of clear identification can stymie reuse as effectively as copyright restrictions – it's not enough to share the resources, organisations also need to share licensing information so that open content can be easily discovered across collections. Sometimes institutional requirements for permission are weighed

upon public domain resources, fostering doubt in place of certainty. Wherever copyright lingers, rights statements bloom in astonishing diversity. The DPLA estimates that there are more than 26,000 *different* rights statements attached to items in their aggregated collection. How are users expected to know what they're allowed to do?

The complexity of copyright fosters confusion and uncertainty beyond the reach of mere law. It's not just legislation that has to change.

Recently Dan Cohen, the Executive Director of the DPLA, argued that the licensing of cultural data should address more than just legal and technical issues. Instead of seeking to enforce acknowledgement of the source of the data through licence conditions, the DPLA wants to push discussion of attribution and reuse 'into the social or ethical realm' by 'pairing a permissive license with a strong moral entreaty'. Instead of a statement about legal constraints, we have the opportunity for a conversation about what matters and why.

It might not figure in our web stats, but the National Library has plenty of anecdotal evidence that Trove changes lives – a grandfather's face glimpsed for the first time, or perhaps a family mystery solved. One man, who grew up in care, found through Trove the only known photograph of himself as a child. How do we weigh such opportunities against questions of ownership and control?

We need to shift the discussion away from the nature of property to the value of use. What do we want to *do*? Online culture is read–write. We do not simply consume

– we share, we remix, we curate and create. Increased participation brings new opportunities for understanding. Do we give them flight or lock them down?

©

One of the main sources of traffic to Trove, up there with Facebook and Wikipedia, is the knitting site Ravelry. Why? Because Ravelry users have found and shared hundreds of craft patterns from Trove's digitised newspapers. And not just shared, but made. The most popular pattern, 'Elegant elephant', discovered in a 1959 edition of the *Australian Women's Weekly*, has been made more than forty times, often with individual embellishments.

This to me seems a great example of the wonderful complexities that the digitisation of cultural heritage collections are introducing into our relationship with the past. A digital version of a fifty-year-old pattern is shared online and spawns a herd of cuddly elephants. From past to present, from digital to physical, the transformations pile one on top of another.

"One of the main sources of traffic to Trove, up there with Facebook and Wikipedia, is the knitting site Ravelry.

It's also unexpected. It's not a use that was designed into the system, it's a new set of experiences brought to life through the passions and ingenuity of Trove users.

The application of digital tools to large cultural collections also promises to surprise. Techniques drawn from computational linguistics, for example, are being used to

analyse the spread of ideas through nineteenth-century newspapers. Another project is using computer vision to identify poems in newspapers through their distinctive shapes. New structures can be found and visualised within digitised sources. New questions can be asked.

But these techniques also carry new risks. Researchers in the digital humanities, exploring the application of technology to fields such as literature and history, have been prominent in discussions around the copyright implications of mass digitisation projects and the legal status of text-mining. Increased clarity around concepts such as 'transformative use' is necessary to ensure that researchers have access to data and the confidence to explore.

And yet the ultimate goal isn't certainty, it's a greater awareness of the constraints around our engagement with the past. Within both government and the cultural sector the value of 'open' data is rightly proclaimed, but open data is always, to some extent, closed. Categories have been assigned, formats have been cleaned, decisions made about what belongs and what doesn't – every spreadsheet contains an argument.

Each elephant is different. The tensions imposed by our overly complex system of copyright do at least remind us that the past can never be 'open', its limits cannot be legislated, its boundaries cannot be fixed. Beneath questions of access and use are better questions about our responsibilities to the past.

©

I'll admit that part of my discomfort in being questioned about the copyright status of 'The real face of White Australia' stemmed from my ignorance. I just didn't know the answer.

I'm still not sure.

I do know now that photographs taken before 1955 are okay. But these were attached to official forms, so perhaps the government owns the copyright. Are they published? I suspect the only way to be certain would be to seek permission from the current government department with responsibility for … what exactly? The White Australia Policy lives on, its workings preserved within the archives.

I'll admit too that I always thought it would be interesting if some part of the government challenged our use of the documents. What exactly would they be claiming ownership of?

'The real face of White Australia' was motivated by a strong sense of responsibility towards those people whose lives are glimpsed through the records. To me the question of responsibility still seems more important than the intricacies of ownership. Our debts are to the people who confront us with their gaze, who defy the legislation that told them they did not belong.

Copyright law will never be able to make these judgements for us. No system can predict the individual ethical calculations that shape our engagement with the past. We may always be confronted with risks.

The word 'access' itself is full of politics. To what? By whom? When it comes to our cultural heritage we should never be satisfied. We must ask about the silences

and the gaps. We must challenge the definitions. Access can never simply be given, to some extent it has to be taken. In the struggle we will find meaning.

There must always be risks. The point is to make the risks count.

Tim Sherratt is a digital historian and cultural data hacker who has been developing online resources relating to libraries, archives, museums and history since 1993. He is currently the Manager of Trove at the National Library of Australia, and Associate Professor of Digital Heritage at the University of Canberra.